

Best Available Copy



4. S. A.
091532533



INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ



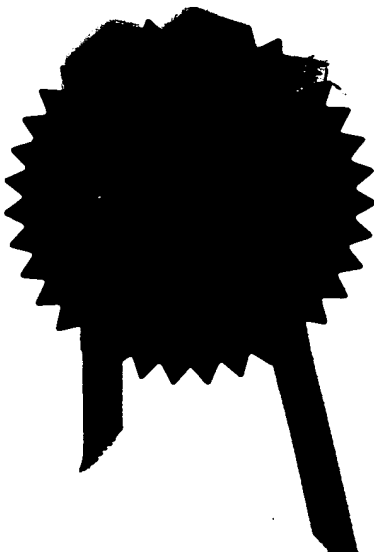
**CERTIFIED COPY OF
PRIORITY DOCUMENT**

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.



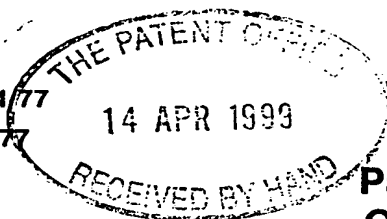
Signed

Dated

30 MAR 2000

THIS PAGE BLANK (USPTO)

Patents Form 1/77
Patents Act 1977
(Rule 16)



The
Patent
Office

15APR99 E439991-6 D02917
P01/7700 0.00 - 9908546.6

Request for grant of a patent

The Patent Office
Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference
2647301
2. Patent Application Number
9908546.6 14 APR 1999
3. Full name, address and postcode of the or of each applicant (underline all surnames)

Canon Kabushiki Kaisha
30-2 3-Chome Shimomaruko
Ohta-Ku
Tokyo
Japan

863010003

Patents ADP number (if known)

If the applicant is a corporate body, give the
country/state of its incorporation

Country: Japan
State:

4. Title of the invention
IMAGE AND SOUND PROCESSING APPARATUS
5. Name of agent
Beresford & Co
"Address for Service" in the United Kingdom
to which all correspondence should be sent
**2/5 Warwick Court
High Holborn
London WC1R 5DJ**
Patents ADP number
6. Priority details
- | | | |
|---------|-----------------------------|----------------|
| Country | Priority application number | Date of filing |
|---------|-----------------------------|----------------|

Patents Form 1/77

7. If this application is divided or otherwise derived from an earlier UK application give details

Number of earlier of application

Date of filing

8. Is a statement of inventorship and or right to grant of a patent required in support of this request?

YES

9. Enter the number of sheets for any of the following items you are filing with this form.

Continuation sheets of this form 0

Description 71

Claim(s) 11

Abstract 1

Drawing(s) 24 + 24 (8)

10. If you are also filing any of the following, state how many against each item.

Priority documents

Translations of priority documents

Statement of inventorship and
right to grant of a patent (*Patents form 7/77*) 1 + 3 copies

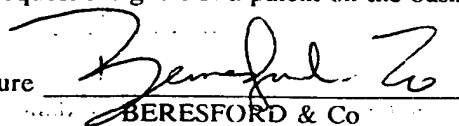
Request for preliminary examination
and search (*Patents Form 9/77*)

Request for Substantive Examination
(*Patents Form 10/77*)

Any other documents
(*please specify*)

11. I/We request the grant of a patent on the basis of this application

Signature


BERESFORD & Co

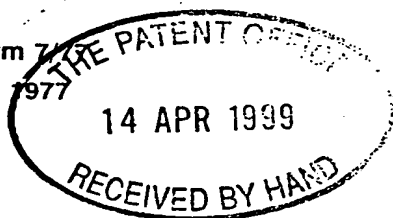
Date 14 April 1999

12. Name and daytime telephone number of
person to contact in the United Kingdom

David SPROSTON

Tei:0171-831-2290

Patents Form 7
Patents Act 1977
(Rule 15)



**The
Patent
Office**

**Statement of inventorship and of
right to grant of a patent**

The Patent Office
Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference
2647301
2. Patent Application Number
accompanying application reference 2647301 99 08546.6
3. Full name of the or each applicant
Canon Kabushiki Kaisha
4. Title of the invention
IMAGE AND SOUND PROCESSING APPARATUS
5. State how the applicant(s) derived the right from the inventor(s) to be granted a patent
By virtue of the employment of the inventors by Canon Research Centre Europe Ltd, and by virtue of an agreement between Canon Research Centre Europe Ltd and Canon Kabushiki Kaisha dated 1 January 1994.
6. How many, if any additional Patents Forms
7/77 are attached to this form?
11. I/We believe that the person(s) named over the page (and on any extra copies of this form) is/are the inventor(s) of the invention which the above patent application relates to.

Signature BERESFORD & Co Date 14 April 1999
12. Name and daytime telephone number of
person to contact in the United Kingdom David SPROSTON
Tel: 0171-831-2290

Patents Form 7/77

TAYLOR; Michael James
c/o CANON RESEARCH CENTRE
EUROPE LTD
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ

RAJAN; Jebu Jacob
c/o CANON RESEARCH CENTRE
EUROPE LTD
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ

ROWE; Simon Michael
c/o CANON RESEARCH CENTRE
EUROPE LTD
1 Occam Court, Occam Road
Surrey Research Park
Guildford
Surrey GU2 5YJ

IMAGE AND SOUND PROCESSING APPARATUS

The present invention relates to the processing of image data and sound data to generate data to assist in
5 archiving the image and sound data.

Many databases exist for the storage of data. However, the existing databases suffer from the problem that the ways in which the database can be interrogated to
10 retrieve information therefrom are limited.

The present invention has been made with this problem in mind.

15 According to the present invention, there is provided an apparatus or method in which image and sound data recording the movements and speech of a number of people is processed using a combination of image processing and sound processing to identify which people shown in the
20 image data are speaking, and sound data is processed to generate text data corresponding to the words spoken using processing parameters selected in dependence upon the identified speaking participant(s).

25 The text data may then be stored in a database together with the image data and/or the sound data to facilitate

information retrieval from the database.

The present invention also provides an apparatus or method in which the positions in three dimensions of a number of people are determined by processing image data, sound data conveying words spoken by the people is processed to determine the direction of the sound source in three dimensions, the speaker of the words is identified using the generated positional information, and voice recognition parameters for performing speech-to-text processing are selected for the identified speaker.

In this way, the speaking participant can be readily identified to enable the sound data to be processed.

Preferably, the position of each person is determined by processing the image data to track at least the head of each person.

20

The present invention further provides an apparatus or method for processing image data and sound data in such a system to identify a speaking participant.

25 The present invention further provides instructions, including in signal and recorded form, for configuring

a programmable processing apparatus to become arranged as an apparatus, or to become operable to perform a method, in such a system.

5 Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

Figure 1 illustrates the recording of sound and video
10 data from a meeting between a plurality of participants;

Figure 2 is a block diagram showing an example of
notional functional components within a processing
apparatus in an embodiment;

15 Figure 3 shows the processing operations performed by processing apparatus 24 in Figure 2 prior to the meeting shown in Figure 1 between the participants starting;

20 Figure 4 schematically illustrates the data stored in meeting archive database 60 at step S2 and step S4 in Figure 3;

Figure 5 shows the processing operations performed at
25 step S34 in Figure 3 and step S70 in Figure 7;

Figure 6 shows the processing operations performed at each of steps S42-1, S42-2 and S42-n in Figure 5;

5 Figure 7 shows the processing operations performed by processing apparatus 24 in Figure 2 while the meeting between the participants is taking place;

Figure 8 shows the processing operations performed at step S72 in Figure 7;

10

Figure 9 shows the processing operations performed at step S80 in Figure 8;

15 Figure 10 illustrates the viewing ray for a participant used in the processing performed at step S114 and step S124 in Figure 9;

Figure 11 illustrates the angles calculated in the processing performed at step S114 in Figure 9;

20

Figure 12 shows the processing operations performed at step S84 in Figure 8;

25 Figure 13 shows the processing operations performed at step S89 in Figure 8;

Figure 14 shows the processing operations performed at step S168 in Figure 13;

Figure 15 schematically illustrates the storage of
5 information in the meeting archive database 60;

Figures 16A and 16B show examples of viewing histograms defined by data stored in the meeting archive database 60;

10

Figure 17 shows the processing operations performed at step S102 in Figure 8;

Figure 18 shows the processing operations performed by
15 processing apparatus 24 to retrieve information from the meeting archive database 60;

Figure 19A shows the information displayed to a user at step S200 in Figure 18;

20

Figure 19B shows an example of information displayed to a user at step S204 in Figure 18; and

Figure 20 schematically illustrates an embodiment in
25 which a single database stores information from a plurality of meetings and is interrogated from one or

more remote apparatus.

Referring to Figure 1, a plurality of video cameras (three in the example shown in Figure 1, although this number may be different) 2-1, 2-2, 2-3 and a microphone array 4 are used to record image data and sound data respectively from a meeting taking place between a group of people 6, 8, 10, 12.

10 The microphone array 4 comprises an array of microphones arranged such that the direction of any incoming sound can be determined, for example as described in GB-A-2140558, US 4333170 and US 3392392.

15 The image data from the video cameras 2-1, 2-2, 2-3 and the sound data from the microphone array 4 is input via cables (not shown) to a computer 20 which processes the received data and stores data in a database to create an archive record of the meeting from which information can
20 be subsequently retrieved.

Computer 20 comprises a conventional personal computer having a processing apparatus 24 containing, in a conventional manner, one or more processors, memory, sound card etc., together with a display device 26 and
25 user input devices, which, in this embodiment, comprise

a keyboard 28 and a mouse 30.

The components of computer 20 and the input and output of data therefrom are schematically shown in Figure 2.

5

Referring to Figure 2, the processing apparatus 24 is programmed to operate in accordance with programming instructions input, for example, as data stored on a data storage medium, such as disk 32, and/or as a signal 34
10 input to the processing apparatus 24, for example from a remote database, by transmission over a communication network (not shown) such as the Internet or by transmission through the atmosphere, and/or entered by a user via a user input device such as keyboard 28 or
15 other input device.

When programmed by the programming instructions, processing apparatus 24 effectively becomes configured into a number of functional units for performing
20 processing operations. Examples of such functional units and their interconnections are shown in Figure 2. The illustrated units and interconnections in Figure 2 are, however, notional and are shown for illustration purposes only, to assist understanding; they do not necessarily
25 represent the exact units and connections into which the processor, memory, etc of the processing apparatus 24

become configured.

Referring to the functional units shown in Figure 2, a central controller 36 processes inputs from the user input devices 28, 30 and receives data input to the processing apparatus 24 by a user as data stored on a storage device, such as disk 38, or as a signal 40 transmitted to the processing apparatus 24. The central controller 36 also provides control and processing for a number of the other functional units. Memory 42 is provided for use by central controller 36 and other functional units.

Head tracker 50 processes the image data received from video cameras 2-1, 2-2, 2-3 to track the position and orientation in three dimensions of the head of each of the participants 6, 8, 10, 12 in the meeting. In this embodiment, to perform this tracking, head tracker 50 uses data defining a three-dimensional computer model of the head of each of the participants and data defining features thereof, which is stored in head model store 52, as will be described below.

Direction processor 53 processes sound data from the microphone array 4 to determine the direction or directions from which the sound recorded by the

microphones was received. Such processing is performed in a conventional manner, for example as described in GB-A-2140558, US 4333170 and US 3392392.

5 Voice recognition processor 54 processes sound data received from microphone array 4 to generate text data therefrom. More particularly, voice recognition processor 54 operates in accordance with a conventional voice recognition program, such as "Dragon Dictate" or
10 IBM "ViaVoice", to generate text data corresponding to the words spoken by the participants 6, 8, 10, 12. To perform the voice recognition processing, voice recognition processor 54 uses data defining the speech recognition parameters for each participant 6, 8, 10, 12,
15 which is stored in speech recognition parameter store 56. More particularly, the data stored in speech recognition parameter store 56 comprises data defining the voice profile of each participant which is generated by training the voice recognition processor in a
20 conventional manner. For example, the data comprises the data stored in the "user files" of Dragon Dictate after training.

Archive processor 58 generates data for storage in
25 meeting archive database 60 using data received from head tracker 50, direction processor 53 and voice recognition

processor 54. More particularly, as will be described below, video data from cameras 2-1, 2-2 and 2-3 and sound data from microphone array 4 is stored in meeting archive database 60 together with text data from voice
5 recognition processor 54 and data defining at whom each participant in the meeting was looking at a given time.

Text searcher 62, in conjunction with central controller 36, is used to search the meeting archive database 60 to
10 find and replay the sound and video data for one or more parts of the meeting which meet search criteria specified by a user, as will be described in further detail below.

Display processor 64 under control of central controller
15 36 displays information to a user via display device 26 and also replays sound and video data stored in meeting archive database 60.

Output processor 66 outputs part or all of the data from
20 archive database 60, for example on a storage device such as disk 68 or as a signal 70.

Before beginning the meeting, it is necessary to initialise computer 20 by entering data which is
25 necessary to enable processing apparatus 24 to perform the required processing operations.

Figure 3 shows the processing operations performed by processing apparatus 24 during this initialisation.

Referring to Figure 3, at step S1, central controller 36
5 causes display processor 64 to display a message on display device 26 requesting the user to input the names of each person who will participate in the meeting.

At step S2, upon receipt of data defining the names, for
10 example input by the user using keyboard 28, central controller 36 allocates a unique identification number to each participant, and stores data, for example table 80 shown in Figure 4, defining the relationship between the identification numbers and the participants' names
15 in the meeting archive database 60.

At step S3, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to input the name of each object at
20 which a person may look for a significant amount of time during the meeting, and for which it is desired to store archive data in the meeting archive database 60. Such objects may include, for example, a flip chart, such as the flip chart 14 shown in Figure 1, a whiteboard or
25 blackboard, or a television, etc.

At step S4, upon receipt of data defining the names of the objects, for example input by the user using keyboard 28, central controller 36 allocates a unique identification number to each object, and stores data, for example as in table 80 shown in Figure 4, defining the relationship between the identification numbers and the names of the objects in the meeting archive database 60.

10 At step S6, central controller 36 searches the head model store 52 to determine whether data defining a head model is already stored for each participant in the meeting.

If it is determined at step S6 that a head model is not already stored for one or more of the participants, then, at step S8, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to input data defining a head model of each participant for whom a model is not already stored.

20 In response, the user enters data, for example on a storage medium such as disk 38 or by downloading the data as a signal 40 from a connected processing apparatus, defining the required head models. Such head models may be generated in a conventional manner, for example as

described in "An Analysis/Synthesis Cooperation for Head Tracking and Video Face Cloning" by Valente et al in Proceedings ECCV '98 Workshop on Perception of Human Action, University of Freiberg, Germany, June 6 1998.

5

At step S10, central controller 36 stores the data input by the user in head model store 52.

At step S12, central controller 36 and display processor
10 64 render each three-dimensional computer head model input by the user to display the model to the user on display device 26, together with a message requesting the user to identify at least seven features in each model.

15 In response, the user designates using mouse 30 points in each model which correspond to prominent features on the front, sides and, if possible, the back, of the participant's head, such as the corners of eyes, nostrils, mouth, ears or features on glasses worn by the
20 participant, etc.

At step S14, data defining the features identified by the user is stored by central controller 36 in head model store 52.

25

On the other hand, if it is determined at step S6 that

a head model is already stored in head model store 52 for each participant, then steps S8 to S14 are omitted.

At step S16, central controller 36 searches speech
5 recognition parameter store 56 to determine whether speech recognition parameters are already stored for each participant.

If it is determined at step S16 that speech recognition
10 parameters are not available for all of the participants, then, at step S18, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to input the speech recognition parameters for each participant for whom the parameters
15 are not already stored.

In response, the user enters data, for example on a storage medium such as disk 38 or as a signal 40 from a remote processing apparatus, defining the necessary
20 speech recognition parameters. As noted above, these parameters define a profile of the user's speech and are generated by training a voice recognition processor in a conventional manner. Thus for example, in the case of a voice recognition processor comprising Dragon Dictate,
25 the speech recognition parameters input by the user correspond to the parameters stored in the "user files"

of Dragon Dictate.

At step S20, data defining the speech recognition parameters input by the user is stored by central controller 36 in the speech recognition parameter store 56.

On the other hand, if it is determined at step S16 that the speech recognition parameters are already available for each of the participants, then steps S18 and S20 are omitted.

At step S22, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to perform steps to enable the cameras 2-1, 2-2 and 2-3 to be calibrated.

In response, the user carries out the necessary steps and, at step S24, central controller 36 performs processing to calibrate the cameras 2-1, 2-2 and 2-3. More particularly, in this embodiment, the steps performed by the user and the processing performed by central controller 36 are carried out in a manner such as that described in Appendix A herewith. This generates calibration data defining the position and orientation of each camera 2-1, 2-2 and 2-3 with respect to the

meeting room and also the intrinsic parameters of each camera (aspect ratio, focal length, principal point, and first order radial distortion coefficient). The camera calibration data is stored, for example in memory 42.

5

At step S25, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to perform steps to enable the position and orientation of each of the objects for which
10 identification data was stored at step S4 to be determined.

In response, the user carries out the necessary steps and, at step S26, central controller 36 performs
15 processing to determine the position and orientation of each object. More particularly, in this embodiment, the user places coloured markers at points on the perimeter of the surface(s) of the object at which the participants in the meeting may look, for example the plane of the
20 sheets of paper of flip chart 14. Image data recorded by each of cameras 2-1, 2-2 and 2-3 is then processed by central controller 36 using the camera calibration data stored at step S24 to determine, in a conventional manner, the position in three-dimensions of each of the
25 coloured markers. This processing is performed for each camera 2-1, 2-2 and 2-3 to give separate estimates of the

position of each coloured marker, and an average is then determined for the position of each marker from the positions calculated using data from each camera 2-1, 2-2 and 2-3. Using the average position of each marker, 5 central controller 36 calculates in a conventional manner the centre of the object surface and a surface normal to define the orientation of the object surface. The determined position and orientation for each object is stored as object calibration data, for example in memory 10 42.

At step S27, central controller 36 causes display processor 64 to display a message on display device 26 requesting the next participant in the meeting (this 15 being the first participant the first time step S27 is performed) to sit down.

At step S28, processing apparatus 24 waits for a predetermined period of time to give the requested 20 participant time to sit down, and then, at step S30, central controller 36 processes the respective image data from each camera 2-1, 2-2 and 2-3 to determine an estimate of the position of the seated participant's head for each camera. More particularly, in this embodiment, 25 central controller 36 carries out processing separately for each camera in a conventional manner to identify each

portion in a frame of image data from the camera which has a colour corresponding to the colour of the skin of the participant (this colour being determined from the data defining the head model of the participant stored in head model store 52), and then selects the portion which corresponds to the highest position in the meeting room (since it is assumed that the head will be the highest skin-coloured part of the body). Using the position of the identified portion in the image and the camera calibration parameters determined at step S24, central controller 36 then determines an estimate of the three-dimensional position of the head in a conventional manner. This processing is performed for each camera 2-1, 2-2 and 2-3 to give a separate head position estimate for each camera.

At step S32, central controller 36 determines an estimate of the orientation of the participant's head in three dimensions for each camera 2-1, 2-2 and 2-3. More particularly, in this embodiment, central controller 36 renders the three-dimensional computer model of the participant's head stored in head model store 52 for a plurality of different orientations of the model to produce a respective two-dimensional image of the model for each orientation. In this embodiment, the computer model of the participant's head is rendered in 108

different orientations to produce 108 respective two-dimensional images, the orientations corresponding to 36 rotations of the head model in 10° steps for each of three head inclinations corresponding to 0° (looking straight ahead), $+45^\circ$ (looking up) and -45° (looking down). Each two-dimensional image of the model is then compared by central processor 36 with the part of the video frame from a camera 2-1, 2-2, 2-3 which shows the participant's head, and the orientation for which the image of the model best matches the video image data is selected, this comparison and selection being performed for each camera to give a head orientation estimate for each camera. When comparing the image data produced by rendering the head model with the video data from a camera, a conventional technique is used, for example as described in "Head Tracking Using a Textured Polygonal Model" by Schödl, Haro & Essa in Proceedings 1998 Workshop on Perceptual User Interfaces.

At step S34, the respective estimates of the position of the participant's head generated at step S30 and the respective estimates of the orientation of the participant's head generated at step S32 are input to head tracker 50 and frames of image data received from each of cameras 2-1, 2-2 and 2-3 are processed to track the head of the participant. More particularly, in this

embodiment, head tracker 50 performs processing to track the head in a conventional manner, for example as described in "An Analysis/Synthesis Cooperation for Head Tracking and Video Face Cloning" by Valente et al in
5 Proceedings EECV '98 Workshop on Perception of Human Action, University of Freiberg, Germany, June 6 1998.

Figure 5 summarises the processing operations performed by head tracker 50 at step S34.

10

Referring to Figure 5, in each of steps S42-1 to S42-n ("n" being three in this embodiment since there are three cameras), head tracker 50 processes image data from a respective one of the cameras recording the meeting to
15 determine the positions of the head features of the participant (stored at step S14) in the image data from the camera and to determine therefrom the three-dimensional position and orientation of the participant's head for the current frame of image data from that
20 camera.

Figure 6 shows the processing operations performed at a given one of steps S42-1 to S42-n, the processing operations being the same at each step but being carried
25 out on image data from a different camera.

Referring to Figure 6, at step S50, head tracker 50 reads the current estimates of the 3D position and orientation of the participant's head, these being the estimates produced at steps S30 and S32 in Figure 3 the first time
5 step S50 is performed.

At step S52, head tracker 50 uses the camera calibration data generated at step S24 to render the three-dimensional computer model of the participant's head
10 stored in head model store 52 in accordance with the estimates of position and orientation read at step S50.

At step S54, head tracker 50 processes the image data for the current frame of video data received from the camera
15 to extract the image data from each area which surrounds the expected position of one of the head features identified by the user and stored at step S14, the expected positions being determined from the estimates read at step S50 and the camera calibration data
20 generated at step S24.

At step S56, head tracker 50 matches the rendered image data generated at step S52 and the camera image data extracted at step S54 to find the camera image data which
25 best matches the rendered head model.

At step S58, head tracker 50 uses the camera image data identified at step S56 which best matches the rendered head model together with the camera calibration data stored at step S24 (Figure 3) to determine the 3D position and orientation of the participant's head for the current frame of video data.

Referring again to Figure 5, at step S44, head tracker 50 uses the camera image data identified at each of steps S42-1 to S42-n which best matches the rendered head model (identified at step S58 in Figure 6) to determine an average 3D position and orientation of the participant's head for the current frame of video data.

At the same time that step S44 is performed, at step S46, the positions of the head features in the camera image data determined at each of steps S42-1 to S42-n (identified at step S58 in Figure 6) are input into a conventional Kalman filter to generate an estimate of the 3D position and orientation of the participant's head for the next frame of video data. Steps S42 to S46 are performed repeatedly for the participant as frames of video data are received from video camera 2-1, 2-2 and 2-3.

25

Referring again to Figure 3, at step S36, central

controller 36 determines whether there is another participant in the meeting, and steps S27 to S36 are repeated until processing has been performed for each participant in the manner described above. However,
5 while these steps are performed for each participant, at step S34, head tracker 50 continues to track the head of each participant who has already sat down.

When it is determined at step S36 that there are no
10 further participants in the meeting and that accordingly the head of each participant is being tracked by head tracker 50, then, at step S38, central controller 36 causes an audible signal to be output from processing apparatus 24 to indicate that the meeting between the
15 participants can begin.

Figure 7 shows the processing operations performed by processing apparatus 24 as the meeting between the participants takes place.

20

Referring to Figure 7, at step S70, head tracker 50 continues to track the head of each participant in the meeting. The processing performed by head tracker 50 at step S70 is the same as that described above with respect
25 to step S34, and accordingly will not be described again here.

At the same time that head tracker 50 is tracking the head of each participant at step S70, at step S72 processing is performed to generate and store data in meeting archive database 60.

5

Figure 8 shows the processing operations performed at step S72.

Referring to Figure 8, at step S80, archive processor 58
10 generates a so-called "viewing parameter" for each participant defining at which person or which object the participant is looking.

Figure 9 shows the processing operations performed at
15 step S80.

Referring to Figure 9, at step S110, archive processor 58 reads the current three-dimensional position of each participant's head from head tracker 50, this being the
20 average position generated in the processing performed by head tracker 50 at step S44 (Figure 5).

At step S112, archive processor 58 reads the current orientation of the head of the next participant (this
25 being the first participant the first time step S112 is performed) from head tracker 50. The orientation read

at step S112 is the average orientation generated in the processing performed by head tracker 50 at step S44 (Figure 5).

5 At step S114, archive processor 58 determines the angle between a ray defining where the participant is looking (a so-called "viewing ray") and each notional line which connects the head of the participant with the centre of the head of another participant.

10

More particularly, referring to Figures 10 and 11, an example of the processing performed at step S114 is illustrated for one of the participants, namely participant 6 in Figure 1. Referring to Figure 10, the orientation of the participant's head read at step S112 defines a viewing ray 90 from a point between the centre of the participant's eyes which is perpendicular to the participant's head. Similarly, referring to Figure 11, the positions of all of the participant's heads read at
15 step S110 define notional lines 92, 94, 96 from the point between the centre of the eyes of participant 6 to the centre of the heads of each of the other participants 8, 10, 12. In the processing performed at step S114, archive processor 58 determines the angles 98, 100, 102
20 between the viewing ray 90 and each of the notional lines
25 92, 94, 96.

Referring again to Figure 9, at step S116, archive processor 58 selects the angle 98, 100 or 102 which has the smallest value. Thus, referring to the example shown in Figure 11, the angle 100 would be selected.

5

At step S118, archive processor 58 determines whether the angle selected at step S116 has a value less than 10° .

If it is determined at step S118 that the angle is less than 10° , then, at step S120, archive processor 58 sets the viewing parameter for the participant to the identification number (allocated at step S2 in Figure 3) of the participant connected by the notional line which makes the smallest angle with the viewing ray. Thus, referring to the example shown in Figure 11, if angle 100 is less than 10° , then the viewing parameter would be set to the identification number of participant 10 since angle 100 is the angle between viewing ray 90 and notional line 94 which connects participant 6 to participant 10.

20

On the other hand, if it is determined at step S118 that the smallest angle is not less than 10° , then, at step S122, archive processor 58 reads the position of each object previously stored at step S26 (Figure 3).

25

At step S124, archive processor 58 determines whether the viewing ray 90 of the participant intersects the plane of any of the objects.

5 If it is determined at step S124 that the viewing ray 90 does intersect the plane of an object, then, at step S126, archive processor 50 sets the viewing parameter for the participant to the identification number (allocated at step S4 in Figure 3) of the object which is
10 intersected by the viewing ray, this being the nearest intersected object to the participant if more than one object is intersected by the viewing ray 90.

On the other hand, if it is determined at step S124 that
15 the viewing ray 90 does not intersect the plane of an object, then, at step S128, archive processor 58 sets the value of the viewing parameter for the participant to "0". This indicates that the participant is determined to be looking at none of the other participants (since
20 the viewing ray 90 is not close enough to any of the notional lines 92, 94, 96) and none of the objects (since the viewing ray 90 does not intersect an object). Such a situation could arise, for example, if the participant was looking at some object in the meeting room for which
25 data had not been stored at step S4 and which had not been calibrated at step S26 (for example the notes held

by participant 12 in the example shown in Figure 1).

At step S130, archive processor 58 determines whether there is another participant in the meeting, and steps
5 S112 to S130 are repeated until the processing described above has been carried out for each of the participants.

Referring again to Figure 8, at step S82, central
controller 36 and voice recognition processor 54
10 determine whether any speech data has been received from the microphone array 4 corresponding to the current frame of video data.

If it is determined at step S82 that speech data has been
15 received, then, at step S84, processing is performed to determine which of the participants in the meeting is speaking.

Figure 12 shows the processing operations performed at
20 step S84.

Referring to Figure 12, at step S140, direction processor
53 processes the sound data from the microphone array 4
to determine the direction or directions from which the
25 speech is coming. This processing is performed in a conventional manner, for example as described in

GB-A-2140558, US 4333170 and US 3392392.

At step S142, archive processor 58 reads the position of each participant's head determined by head tracker 50 at
5 step S44 (Figure 5) for the current frame of image data and determines therefrom which of the participants has a head at a position corresponding to a direction determined at step S140, that is, a direction from which the speech is coming.

10

At step S144, archive processor 58 determines whether there is more than one participant in a direction from which the speech is coming.

15 If it is determined at step S144 that there is only one participant in the direction from which the speech is coming, then, at step S146, archive processor 58 selects the participant in the direction from which the speech is coming as the speaker for the current frame of image
20 data.

On the other hand, if it is determined at step S144 that there is more than one participant having a head at a position which corresponds to the direction from which
25 the speech is coming, then, at step S148, archive processor 58 determines whether one of those participants

was identified as the speaker in the preceding frame of image data.

If it is determined at step S148 that one of the participants in the direction from which the speech is coming was selected as the speaker in the preceding frame of image data, then, at step S150, archive processor 58 selects the speaker identified for the previous frame of image data as the speaker for the current frame of image data, too. This is because is it likely that the speaker in the previous frame of image data is the same as the speaker in the current frame.

On the other hand, if it is determined at step S148 that none of the participants in the direction from which the speech is coming is the participant identified as the speaker in the preceding frame, or if no speaker was identified for the previous frame, then, at step S152, archive processor 58 selects each of the participants in the direction from which the speech is coming as a "potential" speaking participant.

Referring again to Figure 8, at step S86, archive processor 58 stores the viewing parameter value for each speaking participant, that is the viewing parameter value generated at step S80 defining at whom or what each

speaking participant is looking, for subsequent analysis, for example in memory 42.

At step S88, archive processor 58 informs voice
5 recognition processor 54 of the identity of each speaking
participant determined at step S84. In response, voice
recognition processor 54 selects the speech recognition
parameters for the speaking participant(s) from speech
10 recognition parameter store 56 and uses the selected
parameters to perform speech recognition processing on
the received speech data to generate text data
corresponding to the words spoken by the speaking
participant(s).

15 On the other hand, if it is determined at step S82 that
the received sound data does not contain any speech, then
steps S84 to S88 are omitted.

At step S89, archive processor 58 determines which image
20 data is to be stored in the meeting archive database 60,
that is, the image data from which of the cameras 2-1,
2-2 and 2-3 is to be stored.

Figure 13 shows the processing operations performed by
25 archive processor 58 at step S89.

Referring to Figure 13, at step S160, archive processor 58 determines whether any speech was detected at step S82 (Figure 8) for the current frame of image data.

- 5 If it is determined at step S160 that there is no speech for the current frame, then, at step S162, archive processor 58 selects a default camera as the camera from which image data is to be stored. More particularly, in this embodiment, archive processor 58 selects the camera
- 10 or cameras from which image data was recorded for the previous frame, or, if the current frame being processed is the very first frame, then archive processor 58 selects one of the cameras 2-1, 2-2, 2-3 at random.
- 15 On the other hand, if it is determined at step S160 that there is speech for the current frame being processed then, at step S164, archive processor 58 reads the viewing parameter previously stored at step S86 for the next speaking participant (this being the first speaking
- 20 participant the first time step S164 is performed) to determine the person or object at which that speaking participant is looking.

At step S166, archive processor 58 reads the head

25 position and orientation (determined at step S44 in Figure 5) for the speaking participant currently being

considered, together with the head position and orientation of the participant at which the speaking participant is looking (determined at step S44 in Figure 5) or the position and orientation of the object at which the speaking participant is looking (stored at step S26 in Figure 3).

At step S168 archive processor 58 processes the positions and orientations read at step S166 to determine which of the cameras 2-1, 2-2, 2-3 best shows both the speaking participant and the participant or object at which the speaking participant is looking, and selects this camera as a camera from which image data for the current frame is to be stored in meeting archive database 60.

Figure 14 shows the processing operations performed by archive processor 58 at step S168.

Referring to Figure 14, at step S176, archive processor 58 reads the three-dimensional position and viewing direction of the next camera (this being the first camera the first time step S176 is performed), this information having previously been generated and stored at step S24 in Figure 3.

At step S178, archive processor 58 uses the information

read at step S176 together with information defining the
 three-dimensional head position and orientation of the
 speaking participant (determined at step S44 in Figure 5)
 and the three-dimensional head position and orientation
 5 of the participant at whom the speaking participant is
 looking (determined at step S44 in Figure 5) or the
 three-dimensional position and orientation of the object
 being looked at (stored at step S26 in Figure 3) to
 determine whether the speaking participant and the
 10 participant or object at which the speaking participant
 is looking are both within the field of view of the
 camera currently being considered (that is, whether the
 camera currently being considered can see both the
 speaking participant and the participant or object at
 15 which the speaking participant is looking). More
 particularly, in this embodiment, archive processor 58
 evaluates the following equations and determines that the
 camera can see both the speaking participant and the
 participant or object at which the speaking participant
 20 is looking if all of the inequalities hold:

$$\left| \arccos \left[\frac{1}{\sqrt{(X_{p1}-X_c)^2 + (Y_{p1}-Y_c)^2}} \left(\frac{X_{p1}-X_c}{Y_{p1}-Y_c} \cdot \left(\frac{dX_c}{dY_c} \right) \right) \right] \right| < \theta_h$$

....(1)

$$\left| \arccos \left[\frac{1}{\sqrt{(X_{p1}-X_c)^2 + (Y_{p1}-Y_c)^2 + (Z_{p1}-Z_c)^2}} \begin{pmatrix} X_{p1}-X_c \\ Y_{p1}-Y_c \\ Z_{p1}-Z_c \end{pmatrix} \cdot \begin{pmatrix} dX_c \\ dY_c \\ dZ_c \end{pmatrix} \right] \right| < \theta_v$$

....(2)

5

$$\left| \arccos \left[\frac{1}{\sqrt{(X_{p2}-X_c)^2 + (Y_{p2}-Y_c)^2}} \begin{pmatrix} X_{p2}-X_c \\ Y_{p2}-Y_c \end{pmatrix} \cdot \begin{pmatrix} dX_c \\ dY_c \end{pmatrix} \right] \right| < \theta_h$$

....(3)

10

$$\left| \arccos \left[\frac{1}{\sqrt{(X_{p2}-X_c)^2 + (Y_{p2}-Y_c)^2 + (Z_{p2}-Z_c)^2}} \begin{pmatrix} X_{p2}-X_c \\ Y_{p2}-Y_c \\ Z_{p2}-Z_c \end{pmatrix} \cdot \begin{pmatrix} dX_c \\ dY_c \\ dZ_c \end{pmatrix} \right] \right| < \theta_v$$

....(4)

where:

15 (X_c, Y_c, Z_c) are the x, y and z coordinates respectively of the principal point of the camera (previously determined and stored at step S24 in Figure 3)

20 (dX_c, dY_c, dZ_c) represent the viewing direction of the camera in the x, y and z directions respectively (again determined and stored at step S24 in Figure 3)

25 θ_h and θ_v are the angular fields of view of the camera in the horizontal and vertical directions

respectively (again determined and stored at step S24 in Figure 3)

5 (X_{p1}, Y_{p1}, Z_{p1}) are the x, y and z coordinates respectively of the centre of the head of the speaking participant (determined at step S44 in Figure 5)

10 $(dX_{p1}, dY_{p1}, dZ_{p1})$ represent the orientation of the viewing ray 90 of the speaking participant (again determined at step S44 in Figure 5)

15 (X_{p2}, Y_{p2}, Z_{p2}) are the x, y and z coordinates respectively of the centre of the head of the person at whom the speaking participant is looking (determined at step S44 in Figure 5) or of the centre of the surface of the object at which the speaking participant is looking (determined at step S26 in Figure 3)

20 $(dX_{p2}, dY_{p2}, dZ_{p2})$ represent the direction in the x, y and z directions respectively of the viewing ray 90 of the participant at whom the speaking participant is looking (again determined at step S44 in Figure 5) or of the normal to the object

25

surface at which the speaking participant is looking (determined at step S26 in Figure 3).

If it is determined at step S178 that the camera can see both the speaking participant and the participant or object at which the speaking participant is looking (that is, the inequalities in each of equations (1), (2), (3) and (4) above hold), then, at step S180, archive processor 58 calculates and stores a value representing the quality of the view that the camera currently being considered has of the speaking participant. More particularly, in this embodiment, archive processor 58 calculates a quality value, Q1, using the following equation:

$$Q1 = \frac{1}{\sqrt{(X_c - X_{p1})^2 + (Y_c - Y_{p1})^2 + (Z_c - Z_{p1})^2}} \begin{pmatrix} X_c - X_{p1} \\ Y_c - Y_{p1} \\ Z_c - Z_{p1} \end{pmatrix} \cdot \begin{pmatrix} dX_{p1} \\ dY_{p1} \\ dZ_{p1} \end{pmatrix} \quad \dots (5)$$

where the definitions of the terms are the same as those given for equations (1) and (2) above.

20

The quality value, Q1, calculated at step S180 is a scalar, having a value between -1 and +1, with the value being -1 if the back of the speaking participant's head is directly facing the camera, +1 if the face of the speaking participant is directly facing the camera, and

25

a value in-between for other orientations of the speaking participant's head.

At step S182, archive processor 58 calculates and stores
 5 a value representing the quality of the view that the camera currently being considered has of the participant or object at which the speaking participant is looking. More particularly, in this embodiment, archive processor 58 calculates a quality value, Q2, using the following
 10 equation:

$$Q2 = \frac{1}{\sqrt{(X_c - X_{p2})^2 + (Y_c - Y_{p2})^2 + (Z_c - Z_{p2})^2}} \begin{pmatrix} X_c - X_{p2} \\ Y_c - Y_{p2} \\ Z_c - Z_{p2} \end{pmatrix} \cdot \begin{pmatrix} dX_{p2} \\ dY_{p2} \\ dZ_{p2} \end{pmatrix} \dots (6)$$

where the definitions of the parameters are the same as
 15 those given for equations (3) and (4) above.

Again, Q2 is a scalar having a value between -1 if the back of the head of the participant or the back of the surface of the object is directly facing the camera, +1
 20 if the face of the participant or the front surface of the object is directly facing the camera, and values therebetween for other orientations of the participant's head or object surface.

25 At step S184, archive processor 58 compares the quality

value Q1 calculated at step S180 with the quality value Q2 calculated at step S182, and selects the lowest value. This lowest value indicates the "worst view" that the camera has of the speaking participant or the participant or object at which the speaking participant is looking, (the worst view being that of the speaking participant if Q1 is less than Q2, and that of the participant or object at which the speaking participant is looking if Q2 is less than Q1).

10

On the other hand, if it is determined at step S178 that one or more of the equalities in equations (1), (2), (3) and (4) does not hold (that is, the camera can not see both the speaking participant and the participant or object at which the speaking participant is looking), then, steps S180 to S184 are omitted.

15

At step S186, archive processor 58 determines whether there is another camera from which image data has been received. Steps S176 to S186 are repeated until the processing described above has been performed for each camera.

20

At step S188, archive processor 58 compares the "worst view" values stored for each of the cameras when processing was performed at step S184 (that is, the value

25

of Q1 or Q2 stored for each camera at step S184) and selects the highest one of these stored values. This highest value represents the "best worst view" and accordingly, at step S188, archive processor 58 selects
5 the camera for which this "best worst view" value was stored at step S184 as a camera from which image data should be stored in the meeting archive database, since this camera has the best view of both the speaking participant and the participant or object at which the
10 speaking participant is looking.

At step S170, archive processor 58 determines whether there is another speaking participant, including any "potential" speaking participants. Steps S164 to S170
15 are repeated until the processing described above has been performed for each speaking participant and each "potential" speaking participant.

Referring again to Figure 8, at step S90, archive
20 processor 58 encodes the current frame of video data received from the camera or cameras selected at step S89 and the sound data received from microphone array 4 as MPEG 2 data in a conventional manner, and stores the encoded data in meeting archive database 60.

25

Figure 15 schematically illustrates the storage of data

in meeting archive database 60. The storage structure shown in Figure 15 is notional and is provided to assist understanding by illustrating the links between the stored information; it does not necessarily represent the exact way in which data is stored in the memory comprising meeting archive database 60.

Referring to Figure 15, meeting archive database 60 stores time information represented by the horizontal axis 200, on which each unit represents a predetermined amount of time, for example the time period of one frame of video data received from a camera. (It will, of course, be appreciated that the meeting archive database 60 will generally contain many more time units than the number shown in Figure 15.) The MPEG 2 data generated at step S90 is stored as data 202 in meeting archive database 60, together with timing information (this timing information being schematically represented in Figure 15 by the position of the MPEG 2 data 202 along the horizontal axis 200).

Referring again to Figure 8, at step S92, archive processor 58 stores any text data generated by voice recognition processor 54 at step S88 for the current frame in meeting archive database 60 (indicated at 204 in Figure 15). More particularly, the text data is

stored with a link to the corresponding MPEG 2 data, this link being represented in Figure 15 by the text data being stored in the same vertical column as the MPEG 2 data. As will be appreciated, there will not be any text data for storage from participants who are not speaking. In the example shown in Figure 15, text is stored for the first ten time slots for participant 1 (indicated at 206), for the twelfth to twentieth time slots for participant 3 (indicated at 208), and for the twenty-first time slot for participant 4 (indicated at 210). No text is stored for participant 2 since, in this example, participant 2 did not speak during the time slots shown in Figure 15.

At step S94, archive processor 58 stores the viewing parameter value generated for the current frame for each participant at step S80 in the meeting archive database 60 (indicated at 212 in Figure 15). Referring to Figure 15, a viewing parameter value is stored for each participant together with a link to the associated MPEG 2 data 202 and the associated text data 204 (this link being represented in Figure 15 by the viewing parameter values being shown in the same column as the associated MPEG 2 data 202 and associated text data 204). Thus, referring to the first time slot in Figure 15 by way of example, the viewing parameter value for participant 1

is 3, indicating that participant 1 is looking at participant 3, the viewing parameter value for participant 2 is 5, indicating that participant 2 is looking at the flip chart 14, the viewing parameter value for participant 3 is 1, indicating that participant 3 is looking at participant 1, and the viewing parameter value for participant 4 is "0", indicating that participant 4 is not looking at any of the other participants (in the example shown in Figure 1, the participant indicated at 12 is looking at her notes rather than any of the other participants).

At step S96, central controller 36 and archive processor 58 determine whether one of the participants in the meeting has stopped speaking. In this embodiment, this check is performed by examining the text data 204 to determine whether text data for a given participant was present for the previous time slot, but is not present for the current time slot. If this condition is satisfied for any participant (that is, a participant has stopped speaking), then, at step S98, archive processor 58 processes the viewing parameter values previously stored when step S86 was performed for each participant who has stopped speaking (these viewing parameter values defining at whom or what the participant was looking during the period of speech which has now stopped) to

generate data defining a viewing histogram. More particularly, the viewing parameter values for the period in which the participant was speaking are processed to generate data defining the percentage of time during that
5 period that the speaking participant was looking at each of the other participants and objects.

Figures 16A and 16B show the viewing histograms corresponding to the periods of text 206 and 208
10 respectively in Figure 15.

Referring to Figure 15 and Figure 16A, during the period 206 when participant 1 was speaking, he was looking at participant 3 for six of the ten time slots (that is, 60%
15 of the total length of the period for which he was talking), which is indicated at 300 in Figure 16A, and at participant 4 for four of the ten time slots (that is, 40% of the time), which is indicated at 310 in Figure 16A.

20

Similarly, referring to Figure 15 and Figure 16B, during the period 208, participant 3 was looking at participant 1 for approximately 45% of the time, which is indicated at 320 in Figure 16B, at object 5 (that is, the flip
25 chart 14) for approximately 33% of the time, indicated at 330 in Figure 16B, and at participant 2 for

approximately 22% of the time, which is indicated at 340 in Figure 16B.

Referring again to Figure 8, at step S100, each viewing
5 histogram generated at step S98 is stored in the meeting
archive database 60 linked to the associated period of
text for which it was generated. Referring to Figure 15,
the stored viewing histograms are indicated at 214, with
the data defining the histogram for the text period 206
10 indicated at 216, and the data defining the histogram for
the text period 208 indicated at 218. In Figure 15, the
link between the viewing histogram and the associated
text is represented by the viewing histogram being stored
in the same columns as the text data.

15

On the other hand, if it is determined at step S96 that,
for the current time period, one of the participants has
not stopped speaking, then steps S98 and S100 are
omitted.

20

At step S102, archive processor 58 corrects data stored
in the meeting archive database 60 for the previous frame
of video data (that is, the frame preceding the frame for
which data has just been generated and stored at steps
25 S80 to S100) and other preceding frames, if such
correction is necessary.

Figure 17 shows the processing operations performed by archive processor 58 at step S102.

Referring to Figure 17, at step S190, archive processor
5 58 determines whether any data for a "potential" speaking
participant is stored in the meeting archive database 60
for the next preceding frame (this being the frame which
immediately precedes the current frame the first time
step S190 is performed, that is the "i-1"th frame if the
10 current frame is the "i"th frame).

If it is determined at step S190 that no data is stored
for a "potential" speaking participant for the preceding
frame being considered, then it is not necessary to
15 correct any data in the meeting archive database 60.

On the other hand, if it is determined at step S190 that
data for a "potential" speaking participant is stored for
the preceding frame being considered, then, at step S192,
20 archive processor 58 determines whether one of the
"potential" speaking participants for which data was
stored for the preceding frame is the same as a speaking
participant (but not a "potential" speaking participant)
identified for the current frame, that is a speaking
25 participant identified at step S146 in Figure 12.

If it is determined at step S192 that none of the "potential" speaking participants for the preceding frame is the same as a speaking participant identified at step S146 for the current frame, then no correction of the data stored in the meeting archive database 60 for the preceding frame being considered is carried out.

On the other hand, if it is determined at step S192 that a "potential" speaking participant for the preceding frame is the same as a speaking participant identified at step S146 for the current frame, then, at step S194, archive processor 58 deletes the text data 204 for the preceding frame being considered from the meeting archive database 60 for each "potential" speaking participant who is not the same as the speaking participant for the current frame.

By performing the processing at steps S190, S192 and S194 as described above, when a speaker is positively identified by processing image and sound data for the current frame, then data stored for the previous frame for "potential" speaking participants (that is, because it was not possible to unambiguously identify the speaker) is updated using the assumption that the speaker in the current frame is the same as the speaker in the preceding frame.

After step S194 has been performed, steps S190 to S194 are repeated for the next preceding frame. More particularly, if the current frame is the "i"th frame then, the "i-1"th frame is considered the first time steps S190 to S194 are performed, the "i-2"th frame is considered the second time steps S190 to S194 are performed, etc. Steps S190 to S194 continue to be repeated until it is determined at step S190 that data for "potential" speaking participants is not stored in the preceding frame being considered or it is determined at step S192 that none of the "potential" speaking participants in the preceding frame being considered is the same as a speaking participant unambiguously identified for the current frame. In this way, in cases where "potential" speaking participants were identified for a number of successive frames, the data stored in the meeting archive database is corrected if the actual speaking participant from among the "potential" speaking participants is identified in the next frame.

20

Referring again to Figure 8, at step S104, central controller 36 determines whether another frame of video data has been received from the cameras 2-1, 2-2, 2-3. Steps S80 to S104 are repeatedly performed while image data is received from the cameras 2-1, 2-2, 2-3.

25

When data is stored in meeting archive database 60, then the meeting archive database 60 may be interrogated to retrieve data relating to the meeting.

- 5 Figure 18 shows the processing operations performed to search the meeting archive database 60 to retrieve data relating to each part of the meeting which satisfies search criteria specified by a user.
- 10 Referring to Figure 18, at step S200, central controller 36 causes display processor 64 to display a message on display device 26 requesting the user to enter information defining the search of meeting archive database 60 which is required. More particularly, in
- 15 this embodiment, central controller 100 causes the display shown in Figure 19A to appear on display device 26.

- Referring to Figure 19A, the user is requested to enter
- 20 information defining the part or parts of the meeting which he wishes to find in the meeting archive database 60. More particularly, in this embodiment, the user is requested to enter information 400 defining a participant who was talking, information 410 comprising one or more
- 25 key words which were said by the participant identified in information 400, and information 420 defining the

participant or object at which the participant identified in information 400 was looking when he was talking. In addition, the user is able to enter time information defining a portion or portions of the meeting for which the search is to be carried out. More particularly, the user can enter information 430 defining a time in the meeting beyond which the search should be discontinued (that is, the period of the meeting before the specified time should be searched), information 440 defining a time in the meeting after which the search should be carried out, and information 450 and 460 defining a start time and end time respectively between which the search is to be carried out. In this embodiment, information 430, 440, 450 and 460 may be entered either by specifying a time in absolute terms, for example in minutes, or in relative terms by entering a decimal value which indicates a proportion of the total meeting time. For example, entering the value 0.25 as information 430 would restrict the search to the first quarter of the meeting.

20

In this embodiment, the user is not required to enter all of the information 400, 410 and 420 for one search, and instead may omit one or two pieces of this information. If the user enters all of the information 400, 410 and 420, then the search will be carried out to identify each part of the meeting in which the participant identified

25

in information 400 was talking to the participant or object identified in information 420 and spoke the key words defined in information 410. On the other hand, if information 410 is omitted, then a search will be carried out to identify each part of the meeting in which the participant defined in information 400 was talking to the participant or object defined in information 420 irrespective of what was said. If information 410 and 420 is omitted, then a search is carried out to identify each part of the meeting in which the participant defined in information 400 was talking, irrespective of what was said and to whom. If information 400 is omitted, then a search is carried out to identify each part of the meeting in which any of the participants spoke the key words defined in information 410 while looking at the participant or object defined in information 420. If information 400 and 410 is omitted, then a search is carried out to identify each part of the meeting in which any of the participants spoke to the participant or object defined in information 420. If information 420 is omitted, then a search is carried out to identify each part of the meeting in which the participant defined in information 400 spoke the key words defined in information 410, irrespective of to whom the key words were spoken. Similarly, if information 400 and 420 is omitted, then a search is carried out to identify each

part of the meeting in which the key words identified in information 410 were spoken, irrespective of who said the key words and to whom.

- 5 In addition, the user may enter all of the time information 430, 440, 450 and 460 or may omit one or more pieces of this information.

Further, known Boolean operators and search algorithms
10 may be used in combination with key words entered in information 410 to enable the searcher to search for combinations or alternatives of words.

Once the user has entered all of the required information
15 to define the search, he begins the search by clicking on area 470 using a user input device such as the mouse 30.

Referring again to Figure 18, at step S202, the search
20 information entered by the user is read by central controller 36 and the instructed search is carried out. More particularly, in this embodiment, central controller 36 converts any participant or object names entered in information 400 or 420 to identification numbers using
25 the table 80 (Figure 4), and considers the text information 204 for the participant defined in

information 400 (or all participants if information 400 is not entered). If information 420 has been entered by the user, then, for each period of text, central controller 36 checks the data defining the corresponding viewing histogram to determine whether the percentage of viewing time in the histogram for the participant or object defined in information 420 is equal to or above a threshold, which, in this embodiment, is 25%. In this way, periods of speech (text) are considered to satisfy the criteria that a participant defined in information 400 was talking to the participant or object defined in information 420 even if the speaking participant looked at other participants or objects while speaking, provided that the speaking participant looked at the participant or object defined in information 420 for at least 25% of the time of the speech. Thus, for example, a period of speech in which the value of the viewing histogram is equal to or above 25% for two or more participants would be identified if any of these participants were specified in information 420. If the information 410 has been input by the user, then central controller 36 and text searcher 62 search each portion of text previously identified on the basis of information 400 and 420 (or all portions of text if information 400 and 420 was not entered) to identify each portion containing the key word(s) identified in information 410. If any time

information has been entered by the user, then the searches described above are restricted to the meeting times defined by those limits.

5 At step S204, central controller 36 causes display processor 64 to display a list of relevant speeches identified during the search to the user on display device 26. More particularly, central controller 36 causes information such as that shown in Figure 19B to
10 be displayed to the user. Referring to Figure 19B, a list is produced of each speech which satisfies the search parameters, and information is displayed defining the start time for the speech both in absolute terms and as a proportion of the full meeting time. The user is
15 then able to select one of the speeches for playback, for example by clicking on the required speech in the list using the mouse 30.

At step S206, central controller 36 reads the selection
20 made by the user at step S204, and plays back the stored MPEG 2 data 202 for the relevant part of the meeting from meeting archive database 60. More particularly, central controller 36 and display processor 64 decode the MPEG 2 data 202 and output the image data and sound via display
25 device 26. If image data from more than one camera is stored for part, or the whole, of the speech to be played

back, then this is indicated to the user on display device 26 and the user is able to select the image data which is to be replayed by inputting instructions to central controller 36, for example using keyboard 28.

5

At step S208, central controller 36 determines whether the user wishes to cease interrogating the meeting archive database 60 and, if not, steps S200 to S208 are repeated.

10

Various modifications and changes can be made to the embodiment of the invention described above.

In the embodiment above, at step S34 (Figure 3) and step
15 S70 (Figure 7) the head of each of the participants in the meeting is tracked. In addition, however, objects for which data was stored at step S4 and S26 could also be tracked if they moved (such objects may comprise, for example, notes which are likely to be moved by a
20 participant or an object which is to be passed between the participants).

In the embodiment above, image data is processed from a plurality of video cameras 2-1, 2-2, 2-3. However,
25 instead, image data may be processed from a single video camera. In this case, for example, only step S42-1

(Figure 5) is performed and steps S42-2 to S42-n are omitted. Similarly, step S44 is omitted and the 3D position and orientation of the participant's head for the current frame of image data are taken to be the 3D position and orientation determined at step S58 (Figure 6) during the processing performed at step S42-1. At step S46, the position for the head features input to the Kalman filter would be the position in the image data from the single camera. Further, step S89 (Figure 8) to select the camera from which image data is to be recorded in the meeting archive database 60 would also be omitted.

In the embodiment above, at step S168 (Figure 13), processing is performed to identify the camera which has the best view of the speaking participant and also the participant or object at which the speaking participant is looking. However, instead of identifying the camera in the way described in the embodiment above, it is possible for a user to define during the initialisation of processing apparatus 24 which of the cameras 2-1, 2-2, 2-3 has the best view of each respective pair of the seating positions around the meeting table and/or the best view of each respective seating position and a given object (such as flip chart 14). In this way, if it is determined that the speaking participant and the participant at whom the speaking participant is looking

are in predefined seating positions, then the camera defined by the user to have the best view of those predefined seating positions can be selected as a camera from which image data is to be stored. Similarly, if the speaking participant is in a predefined position and is looking at an object, then the camera defined by the user to have the best view of that predefined seating position and object can be selected as the camera from which image data is to be stored.

10

In the embodiment above, at step S162 (Figure 13) a default camera is selected as a camera from which image data was stored for the previous frame. Instead, however, the default camera may be selected by a user, for example during the initialisation of processing apparatus 24.

15

In the embodiment above, at step S194 (Figure 17), the text data 204 is deleted from meeting archive database 60 for the "potential" speaking participants who have now been identified as actually not being speaking participants. In addition, however, the associated viewing histogram data 214 may also be deleted. Further, if MPEG 2 data 202 from more than one of the cameras 2-1, 2-2, 2-3 was stored, then the MPEG 2 data related to the "potential" speaking participants may also be deleted.

20

25

In the embodiment above, when it is not possible to uniquely identify a speaking participant, "potential" speaking participants are defined, data is processed and stored in meeting archive database 60 for the potential speaking participants, and subsequently the data stored in the meeting archive database 60 is corrected (step S102 in Figure 8). However, instead, rather than processing and storing data for "potential" speaking participants, video data received from cameras 2-1, 2-2 and 2-3 and audio data received from microphone array 4 may be stored for subsequent processing and archiving when the speaking participant has been identified from data relating to a future frame. Alternatively, when the processing performed at step S144 (Figure 12) results in an indication that there is more than one participant in the direction from which the speech is coming, image data from the cameras 2-1, 2-2 and 2-3 may be processed to detect lip movements of the participants and to select as the speaking participant the participant in the direction from which the speech is coming whose lips are moving.

In the embodiment above, processing is performed to determine the position of each person's head, the orientation of each person's head and a viewing parameter for each person defining at whom or what the person is

looking. The viewing parameter value for each person is then stored in the meeting archive database 60 for each frame of image data. However, it is not necessary to determine a viewing parameter for all of the people. For example, it is possible to determine a viewing parameter for just the speaking participant, and to store just this viewing parameter value in the meeting archive database 60 for each frame of image data. Accordingly, in this case, it would be necessary to determine the orientation of only the speaking participant's head. In this way, processing requirements and storage requirements can be reduced.

In the embodiment above, at step S202 (Figure 18), the viewing histogram for a particular portion of text is considered and it is determined that the participant was talking to a further participant or object if the percentage of gaze time for the further participant or object in the viewing histogram is equal to or above a predetermined threshold. Instead, however, rather than using a threshold, the participant or object at whom the speaking participant was looking during the period of text (speech) may be defined to be the participant or object having the highest percentage gaze value in the viewing histogram (for example participant 3 in Figure 16A, and participant 1 in Figure 16B).

In the embodiment above, the MPEG 2 data 202, the text data 204, the viewing parameters 212 and the viewing histograms 214 are stored in meeting archive database 60 in real time as data is received from cameras 2-1, 2-2 and 2-3 and microphone array 4. However, instead, the video and sound data may be stored and data 202, 204, 212 and 214 generated and stored in meeting archive database 60 in non-real-time.

10 In the embodiment above, the MPEG 2 data 202, the text data 204, the viewing parameters 212 and the viewing histograms 214 are generated and stored in the meeting archive database 60 before the database is interrogated to retrieve data for a defined part of the meeting.

15 However, some, or all, of the viewing histogram data 214 may be generated in response to a search of the meeting archive database 60 being requested by the user by processing the data already stored in meeting archive database 60, rather than being generated and stored prior

20 to such a request. For example, although in the embodiment above the viewing histograms 214 are calculated and stored in real-time at steps S98 and S100 (Figure 8), these histograms could be calculated in response to a search request being input by the user.

25

In the embodiment above, text data 204 is stored in

meeting archive database 60. Instead, audio data may be stored in the meeting archive database 60 instead of the text data 204. The stored audio data would then either itself be searched for key words using voice recognition processing or converted to text using voice recognition processing and the text search using a conventional text searcher.

In the embodiment above, processing apparatus 24 includes functional components for receiving and generating data to be archived (for example, central controller 36, head tracker 50, head model store 52, direction processor 53, voice recognition processor 54, speech recognition parameter store 56 and archive processor 58), functional components for storing the archive data (for example meeting archive database 60), and also functional components for searching the database and retrieving information therefrom (for example central controller 36 and text searcher 62). However, these functional components may be provided in separate apparatus. For example, one or more apparatus for generating data to be archived, and one or more apparatus for database searching may be connected to one or more databases via a network, such as the Internet.

25

Also, referring to Figure 20, video and sound data from

one or more meetings 500, 510, 520 may be input to a data processing and database storage apparatus 530 (which comprises functional components to generate and store the archive data), and one or more database interrogation
5 apparatus 540, 550 may be connected to the data processing and database storage apparatus 530 for interrogating the database to retrieve information therefrom.

10 In the embodiment above, processing is performed by a computer using processing routines defined by programming instructions. However, some, or all, of the processing could be performed using hardware.

15 Although the embodiment above is described with respect to a meeting taking place between a number of participants, the invention is not limited to this application, and, instead, can be used for other applications, such as to process image and sound data on
20 a film set etc.

Different combinations of the above modifications are, of course, possible and other changes and modifications can be made without departing from the spirit and scope
25 of the invention.

The contents of the applicant's co-pending UK applications 9905191.4, 9905197.1, 9905202.9, 9905158.3, 9905201.1, 9905186.4, 9905160.9, 9905199.7 and 9905187.2 are hereby incorporated by reference.

APPENDIX A

Calibrating and 3D modelling with a multi-camera system

Charles Wiles and Allan Davison

Computer Vision Group, Canon Research Centre Europe
Guildford, Surrey, UK, GU2 5YJ

Abstract

This paper describes a simple and novel way for calibrating the position and internal camera parameters of a camera viewing a scene with no prior knowledge of the camera being necessary. Only two views of a simple planar grid of spots are used to accurately determine the relative position of each camera in a multiple camera system. A multiple camera system is necessary for modelling dynamic objects (such as people). When the shape of the object is continually changing a large number of images must be taken simultaneously. The multiple camera system is also an important research tool allowing surface generation algorithms to be investigated under known accuracy in the camera positions. We have evaluated our algorithm's performance using simulations to determine the limits on the accuracy of our system and have demonstrated the performance in practice by producing 3D models from a four camera system.

1 Introduction

1.1 Motivation

Computing 3D models of a scene from multiple images observing the scene involves two key steps. First the relative position of the camera to the object being modelled must be determined for each image (*camera solving*), second the 3D structure of the object is computed by intersecting the coloured rays observed in the pixels of each image (*surface generation*).

Various methods exist for computing camera positions. When a single hand-held camera is used to record multiple images of a static scene from different positions the position of the camera can be computed by matching distinguishable *features* on surfaces in the scene between views and employing a *structure from motion* algorithm. Although such an approach works well when the features are accurately matched it can fail when few distinguishable features are visible in the scene. Moreover, such a system fails when the scene is dynamic, containing for example a human being.

To avoid feature matching problems prior to camera

solving the camera positions can be either computed by observing a *calibration* object in the scene or *measured* directly using an alternative device. To model an arbitrary dynamic scene it is necessary to record multiple images from different views at the same instant in time; hence multiple cameras are necessary.

For these reasons, we have explored the use of a calibrated multi-camera 3D modelling system. Not only does such a system allow dynamic objects with few distinguishable features to be modelled, but it provides a valuable research tool for investigating surface generation algorithms, since the accuracy of the camera positions can be independently established. Indeed, the accuracy, coverage and robustness plus the choice of algorithm for surface generation depends greatly on the accuracy of the camera positions determined.

1.2 Issues

There are several important research issues concerning calibrated multiple camera systems:

- Prior knowledge of camera intrinsic parameters
- Ease of production of calibration object
- Ease of calibration process
- The number of images that need to be taken (few images are suitable for still cameras, whereas many images can be used for video cameras).
- Range of camera positions from which the camera can view the calibration object well enough to be calibrated
- Accuracy of calibration
- Accuracy of matching image data to the calibration model

There is clearly a trade off between such factors since the most accurate calibration process would likely require a complicated calibration object and process. However, one of the key aims of our work

has been to find the simplest object and process possible to give us a predefined accuracy in calibration.

We define *object space* to be the limit on 3D space within which the object to be modelled is assumed to lie (for a human being this might be a 2x2x2m cube of space). Our goal in calibrating the camera positions is to recover them such that any point within object space projects to within 1 pixel of its true projection in the image. By achieving *maximum projection error* of less than 1 pixel we limit the range of search for agreeing texture in image co-ordinates to 1 pixel from predicted positions during surface generation.

Unfortunately this does not give us a clear measure of the accuracy of the surface generated when correct matches are found between images under surface generation. It does, however, guarantee that the surface of the model will project to within 1 pixel of its true observed position in the original images. We argue that this measure of accuracy on camera position is more important when the goal is to generate a model that is *photo-consistent* with the original images.

1.3 Background

The most accurate calibration object and process would be to have a known 3D point observed in the image for every point in object space and to compute the transfer equation that projects these points into the observed image co-ordinates.

In practice the act of projection is assumed to be a simple parameterised type called a *camera model*. If the parameters (known as the *intrinsic parameters*) of this projection are known then only three world-to-image point matches are required in order to fix the six degrees of freedom in the unknown orientation and location of the camera (known as the camera *extrinsic parameters*)¹. If the intrinsic parameters are not known then, depending on the nature of the calibration object, it is possible to compute these parameters at the same time as the extrinsic parameters from a larger number of matches². In practice noise in the image measurements of the observed points is inevitable and many matches are required so that the maximum likelihood solution can be found by least squares.

In our work we have assumed that the camera model may radially distort the image during projection and that the intrinsic parameters are unknown in advance. Given this starting point the problem is significantly more complicated than computing just

the extrinsic parameters when the intrinsic parameters are known in advance. One reason for taking this approach is that although manufacturers often provide accurate values for the *focal length* of their cameras, the position of the *principal point* often varies greatly and is unknown. Accurate knowledge of the principal point is vital for computing accurate camera position from coplanar world points (see [4]).

There are several methods that have been used to calibrate both intrinsic and extrinsic parameters of a camera. Perhaps the simplest method is to take a single image of a planar calibration grid of known structure. This method was pioneered by Tsai [7]. Although this method is simple, it is only capable of completely calibrating both the intrinsic and extrinsic parameters of the camera if the imaging process exhibits significant radial distortion. If there is little or no radial distortion, the position of the principal point in the image cannot be determined independently from the height of the camera above the calibration plane, and hence calibration fails to provide a complete answer.

In order to calibrate the intrinsic parameters from a planar grid when no radial distortion occurs, multiple images of the grid must be taken. This is the method used by Kanade et al [2] to calibrate their multi-camera rig for dynamic 3D modelling of people. First a planar grid is moved randomly around in front of each video camera to calibrate the intrinsic parameters of the camera, second each camera's position is computed relative to a grid of spots on the floor. The method gives good accuracy of calibration, but uses many images for each camera to carry out the calibration. Our method, presented later, adapts this method to work with just two images from a still camera.

The problem with a single image of a planar grid is solved if an accurate three-dimensional calibration grid is manufactured. Typically an "L" shaped grid is used which has a pattern of black squares on a white background on each of the two flat surfaces. Such calibration grids allow reasonably accurate calibration of the camera to be performed from a single image of the grid and this method has been used extensively for calibrating stereo-camera rigs [1]. The main drawback of such an approach is that accurate manufacturing of the calibration grid is necessary and this is both awkward and expensive. Moreover, such a system does not scale well to large environments and the calibration object must be carefully orientated so that all cameras see a "good" view of the object.

An alternative method is to use a planar calibration

¹In fact, exactly four different solutions are obtained under projective imaging conditions and exactly two different solutions are obtained under affine imaging conditions.

²For example for the perspective camera model, the four intrinsic and six extrinsic parameters can be computed from five 2D-3D point matches when the points are non-coplanar.

grid, but to move it through space in a *known* motion. Either a robot arm or a stepper motor is used to move the grid. By moving the grid in known steps and taking an image after each step, the whole object space can be swept out providing a wealth of matches. Such systems can provide very accurate camera calibration, and although the planar calibration object can be easily made the main drawback is in the expense of the robot arm or stepper motor and in the lack of scalability of the system. Since many images should be recorded for each camera, it is suitable for calibrating video cameras, but is not an ideal method for calibrating still cameras.

More recently "magic wands" [10] have been used to calibrate multiple camera systems. The magic wand approach is to move a single point (or pair of points one at each end of a wand) by hand randomly around in space. Typically the scene is darkened and a point light source is used. The actual algorithm for computing the extrinsic parameters is similar to self-calibration [6] from an unknown object (or structure from motion, if the intrinsic parameters of the cameras are known), but is simpler since the matching problem is trivially solved between images. This method scales well and is inexpensive. It is an excellent method for calibrating multiple video camera rigs, but due to only one point appearing in each image it is not suitable for the calibration of multiple still camera rigs. A further consideration for the magic wand approach are that if a full self-calibration technique is used with a wand with a single point certain camera configurations must be avoided (for example, the cameras must not all lie in the same plane). However if a wand with points at both ends is used then the ambiguity in the Euclidean reconstruction is removed and full self-calibration is possible.

2 Accurate calibration from a planar grid

Our method for calibrating a multiple still camera rig uses a simple planar calibration grid of regularly spaced black circles on a white background. Such a grid is trivial to manufacture (by for example printing on a standard home printer).

The challenge then is to accurately calibrate the position and intrinsic parameters of a number of cameras that observe the grid. Our method stems from the observation that all the intrinsic parameters of a camera can be accurately determined from a small number of images of a planar grid taken with the camera at different positions without needing to know any of the camera positions in advance. Indeed if the camera positions are chosen reasonably carefully, the calibration

can be done from just two images. If one of these images was recorded with the camera in its final position in the multi-camera system then accurate calibration of the multi-camera system can be achieved with just two images from each camera. Additional images provide greater accuracy to the calibration.

The process for calibration is as follows:

1. Take an image of the calibration grid with the camera rotated by roughly 90° about the viewing direction and with a different tilt from the final orientation in the multiple camera rig.
2. Put the camera in its final position in the multiple camera rig and secure in place. Take a second image of the calibration grid.

With the camera in its final position, the calibration grid can be removed and objects to be modelled placed within the space observed by the multi-camera rig.

2.1 Calibrating from a single image

Tsai [7] pioneered the process of calibrating a camera from a planar grid. Tsai's camera model projecting world point $X = (X, Y, Z)^T$ to image point $x = (x, y)^T$ is defined by the set of equations:

$$x = 1/(1 + \kappa_1 R_n^2) \begin{bmatrix} \xi f & 0 \\ 0 & f \end{bmatrix} x_n + x_0,$$

where

$$R_n^2 = x_n^2 + y_n^2$$

and

$$x_n = \frac{1}{Z_c} \begin{pmatrix} X_c \\ Y_c \end{pmatrix}, \quad \text{where } X_c = RX + t.$$

R and t are the 3×3 rotation matrix and translation vector representing the position of the camera, and aspect ratio ξ , focal length f , principal point $x_0 = (x_0, y_0)^T$ and first order radial distortion coefficient κ_1 are the five intrinsic parameters of the camera. If the aspect ratio is known and the distortion coefficient is significant then all the intrinsic and extrinsic parameters can be computed from five or more world-to-image point matches. As well as the radial distortion coefficient being significant it is important that the projection of the grid in the image exhibits significant perspective effects. Hence the camera cannot be calibrated if it is far from the calibration grid or if the viewing direction is close to perpendicular to the plane of the grid. In practice this is easy to avoid.

However, when the first order radial distortion coefficient is zero then the imaging equation reduces to the standard, linear perspective camera model which

has 10 parameters (4 intrinsic + 6 extrinsic). In this case the planar grid to image mapping is completely defined by a linear planar homography containing 8 independent parameters. Hence even when the aspect ratio is known the total number of parameters to be estimated in order to calibrate the camera is still 9 (3 unknown intrinsic + 6 extrinsic) which is clearly not possible.

In the rest of this section we argue that the position of the principal point in the direction parallel to the plane of the calibration grid is determined uniquely, but that the component perpendicular to the plane of the calibration grid cannot be determined independently from the rest of the camera parameters. Indeed we gain an insight into this problem by considering the case when the x -axis of the camera is known to be parallel to the XZ -plane of the world so that the roll of the camera is zero with respect to this plane.

The equation for the perspective camera model in homogeneous co-ordinates is

$$\begin{pmatrix} x \\ 1 \end{pmatrix} = P \begin{pmatrix} X \\ 1 \end{pmatrix},$$

where

$$P = K [R \quad t] \quad , \quad K = \begin{bmatrix} \xi f & 0 & x_0 \\ 0 & f & y_0 \\ 0 & 0 & 1 \end{bmatrix}$$

and equality is only defined up to an arbitrary scale factor. In this case of no camera roll we can rewrite the perspective camera model in terms of only the pitch, α , and yaw, β , of the rotation matrix:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_\alpha & s_\alpha \\ 0 & -s_\alpha & c_\alpha \end{bmatrix} \begin{bmatrix} c_\beta & 0 & s_\beta \\ 0 & 1 & 0 \\ -s_\beta & 0 & c_\beta \end{bmatrix}$$

where c_θ and s_θ are the cosine and sine of angle θ respectively.

If we further note that under calibration from a planar grid, the world points on the grid are considered to be in the world plane $Y = 0$, then we can remove the second column of the rotation matrix, R_2 , and the Y structure co-ordinate from the equation for projection. Thus we can rewrite the equation for projection in terms of the 3x3 planar homography H ,

$$\begin{pmatrix} x \\ 1 \end{pmatrix} = K [R_1 \quad R_3 \quad t] \begin{pmatrix} X \\ Z \\ 1 \end{pmatrix} = H \begin{pmatrix} X \\ Z \\ 1 \end{pmatrix}.$$

Simplification of H leads to,

$$\begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{bmatrix} f_g c_\beta - x_0 s_\beta & f_g s_\beta - x_0 c_\beta & t'_x \\ y_h (-s_\beta) & y_h c_\beta & t'_y \\ -s_\beta & c_\beta & t'_z \end{bmatrix} \begin{bmatrix} X \\ Z \\ 1 \end{bmatrix}$$

where $f_g = \xi f / c_\alpha$, $y_h = y_0 + f \tan(\alpha)$ is the *horizon* of the calibration plane in the image and t' contains some reparameterisation of the vector t .

The important observation is that only seven independent parameters can be observed in H despite having nine elements in the matrix (we have fixed the scale using the bottom row of the matrix). Having fixed the roll angle to be zero and assuming that the aspect ratio is known, we would like to compute all eight unknown perspective camera parameters (3 unknown intrinsic + 5 unknown extrinsic). However of these only the yaw, and the x co-ordinate of the principal point can actually be computed.

Although the argument above is valid for only special case of zero camera roll we have observed empirically that for arbitrary camera orientation, the roll, yaw and component of the principal point parallel to the plane of the calibration grid can indeed be observed from a single image for arbitrary camera roll. The pitch, focal length, translation vector and component of the principal point perpendicular to the plane of the calibration grid cannot be observed.

2.2 Calibrating from a pair of images

The observation that for a perspective camera one component of the principal point can be determined from a single view, but that the perpendicular component cannot, leads directly to a simple method for calibration from two views of the calibration grid.

With two views of the grid, assuming the intrinsic parameters of the camera are unchanged between views, there are 16 independent parameters determined by the two homographies and there are 16 parameters to be estimated in the two cameras (4 fixed intrinsic parameters and 2 sets of 6 extrinsic parameters). With known aspect ratio there are only 15 parameters to be estimated, but in either case there are theoretically enough equations to determine all the unknown parameters.

The key question then is, given that there are theoretically enough equations to solve for all the unknowns, *under what conditions is this possible?*

Recall that in the special case of zero roll we are left with two equations in the unobservable parameters:

$$f_g = \xi f / c_\alpha \quad \text{and} \quad y_h = y_0 + f \tan(\alpha).$$

We will consider that we require only to determine the four remaining unknown parameters in these two equations having solved from a single view for the other intrinsic and extrinsic parameters. Each additional view of the calibration grid gives one more set of these two equations but introduces one additional unknown (the pitch of the camera in the new view).

For the case of known aspect ratio and two views then the equations above give four equations in four remaining unknown parameters. Examination of the equations above show that as long as the pitch of the camera changes between the two views of the calibration grid, then the camera can be fully calibrated. Our first conjecture is then

- if the aspect ratio is known in advance then as long as the pitch of the camera changes between two views then the camera can be fully calibrated from two views of a planar calibration grid.

For the case of unknown aspect ratio there are four equations in five remaining unknown parameters and four equations. Hence calibration cannot be computed from two views of the grid. Our second conjecture is then

- if the aspect ratio is not known in advance and there is no change in the roll angle between the two views then the camera cannot be fully calibrated from two views of a planar calibration grid.

However, for the case of when there is a change in the roll angle of the camera between two views then an interesting simplification occurs, since the principal point becomes uniquely determined. This is because each view fixes the principal point to lie on a specific line in the image, the line being perpendicular to the plane of the calibration grid. If the roll angle changes these lines will not be parallel and hence will intersect at the location of the principal point. Algebraically, y_0 is effectively known in the above equations reducing the number of unknown parameters from five to four. So our third and final conjecture is that:

- if the aspect ratio is not known in advance and there is both a change in the roll angle and a change in the tilt angle between the two views then the camera can be fully calibrated from two views of a planar calibration grid.

Hence in order for the camera intrinsic parameters to be determined fully from two views there must be relative roll (with respect to the plane of the calibration grid) between the positions of the two cameras. Indeed ideally the relative roll should be 90° .

2.3 Accurate feature location

Our approach to matching the calibration object to image data has been to extract the location in the image of point features of known location on the calibration object forming world-to-image point matches. These co-ordinates are then fed into an algorithm for calibrating the camera parameters. A key goal of our

work is to determine the accuracy in feature location needed in order to guarantee a maximum projection error in object space of less than 1 pixel.

More accurate calibration may well be achievable by following the initial "calibration from matches" stage with an iterative "template matching" stage. This second stage would aim to globally minimise the difference between a template of the calibration object and the raw image data. We have not yet implemented such a stage and hence our evaluation for calibration performance is based on the first stage only and may be improved by template matching.

3 Results

3.1 Synthetic data

In order to validate experimentally our prediction that only one component of the principal point can be localised accurately under calibration from a single plane we performed the following experiment using synthetic data.

First the intrinsic parameters and position of a camera were computed fully automatically from a typical view of a calibration grid with the aspect ratio known in advance. The roll of the camera with respect to the plane of the calibration grid was zero and the camera was tilted so that the centre of the grid projects to the centre of the image. The grid fills the unit square and is centred at $(0.5, 0.5, 0.0)$ in the world. The camera is at $(0.7, -1.0, 1.0)$. Since the roll of the camera is zero we expect that the x co-ordinate of the principal point will be accurately located and the y co-ordinate less accurately located getting progressively worse as the amount of radial distortion decreases.

We then repeated each experiment using two views of the grid, one view as in the single view experiment and a second view rotated by 90° about the viewing direction. The aspect ratio was calculated being assumed unknown in advance.

In both cases the maximum projection error within the unit cube of space above the calibration grid was measured as well as the projection error of the point in the centre of the unit cube.

Tsai's freely available code was used to carry out the minimisation with slight modification to enable minimisation of the two view case. Note that for the case of a single view with no radial distortion the minimisation is clearly under-constrained and hence it would be wise reparameterise the solution. However we found this to be unnecessary since the full minimisation reliably and rapidly converged on a solution in which the observable parameters were accurately recovered.

The process for solving for the two view case was as follows:

1. Solve for each view separately using Tsai's single view approach (using a fixed prior estimate for the aspect ratio).
2. Determine the roll of the camera.
3. From the principal point for each view extract the component parallel to the calibration plane and combine into an initial estimate of the principal point.
4. Solve again for each view separately keeping this principal point fixed.
5. Average the intrinsic parameters from each view and use these and the two sets of extrinsic parameters as an initial estimate for a final full two view minimisation.

These experiments were then repeated with non zero roll. The solution computed from a single view accurately recovers the roll of the camera in the extrinsic parameters and hence it is straight forward to extract and combine the components of the principal point parallel to the calibration plane as a precursor to full two view minimisation.

3.1.1 Effect of varying RMS image noise on errors

In this experiment the performance of the full calibration method was measured as the noise in the image co-ordinates of the matches was varied. The experiment was repeated with varying amounts of radial distortion. The full results can be found in [8].

The main observations for calibration from a single view were:

- The estimation of the y co-ordinate of the principal point, y_0 , is computed less accurately than the x co-ordinate of the principal point, x_0 .
- The accuracy of the y co-ordinate of the principal point decreases as the amount of radial distortion decreases.
- The accuracy of the x co-ordinate of the principal point is independent of the amount of radial distortion.
- The maximum projection error of the unit cube sitting directly above the calibration grid is highly correlated with the error in y_0 .

- The maximum projection error occurs for points farthest from the calibration grid and is approximately ten times greater than the error in the projection of the centre point of the unit cube.

In essence when the roll of the camera is zero the estimation of the y_0 intrinsic parameter from a single view is poorly constrained. This leads to errors in projection of 3D points and the further away from the plane of the calibration grid the 3D point being projected is the greater the error in the projection. However, the x_0 intrinsic parameter is well constrained.

The main observations from two views were:

- The error in the y_0 intrinsic parameter was recovered to the same level of accuracy as the x_0 parameter irrespective of the amount of radial distortion
- The maximum projection error for both views was greatly reduced compared with the projection error observed from a single view with the same value of image noise. This difference became more marked the lower the amount of radial distortion.
- With no radial distortion in order to calibrate the camera so that the maximum projection error is guaranteed to be less than one pixel the RMS image noise in the image co-ordinates of features must be less than 0.05 pixels.
- With no radial distortion in order to guarantee the centre projection error to be less than one pixel the image noise must be less than 0.25 pixels.

3.2 Real data

Two experiments were carried out, both using the same set of four off-the-shelf PowershotA5 cameras. In the first experiment, the cameras were positioned around a toy dinosaur to demonstrate small scale modelling, whereas in the second experiment the cameras were positioned in a room so that a person could be modelled.

3.3 Toy dinosaur

A calibration grid printed onto a sheet of A4 paper was placed on a stand. Each of the four cameras were held in a portrait orientation and a photograph of the grid taken as shown in the first column of Figure 2.

Then the cameras were set up as shown in Figure 1 with each camera in a landscape orientation. Three cameras were positioned roughly by hand so that they were about 20cm above the plane of the calibration grid and evenly spaced in a circle about the grid. The fourth camera was positioned so that it was roughly

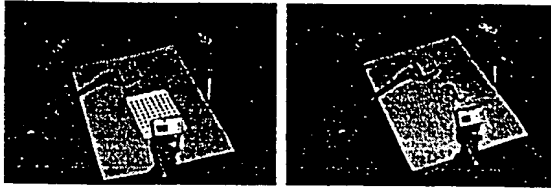


Figure 1: Camera configuration. a) calibration grid imaged, b) toy dinosaur imaged.

above the centre of the grid looking down. A second photograph of the grid was taken with each camera as shown in the second column of Figure 2.

The grid was then removed and in keeping with the indomitable spirit of the academic computer vision community a toy dinosaur placed in the space where the grid had been. A photograph of the dinosaur was taken with each camera.

The three images from each camera were then downloaded to a PC. Two frame calibration as described above was carried out for each camera in turn to calculate its intrinsic parameters and position with respect to the calibration grid in the second image.

Finally the dinosaur images were segmented from the background using a blue screening technique and a voxel carve applied to work out an outer bound on the space occupied by the toy dinosaur. The resulting voxelisation was transformed using a marching cubes algorithm into a faceted VRML model for display. Figure 3 shows the resulting model viewed from the same direction as given in one of the images. The accuracy of the camera calibration is demonstrated by a plausible reconstruction of the toy dinosaur. In particular the tail of the dinosaur is well reconstructed.

3.3.1 Person

A calibration grid was made by sticking together 63 sheets of A4 paper each with a single black circle printed on each. The 7x9 grid occupied a space approximately 1.5m x 1.5m.

First a photograph of the grid was taken with each camera in a landscape orientation. Then the cameras were placed in portrait orientation so that they were roughly evenly spaced through 180°. A second photograph was taken of the calibration grid as shown in the first row of Figure 4. Finally the calibration grid was removed from the scene and a photograph of a person taken with each camera.

A faceted model was reconstructed as before as

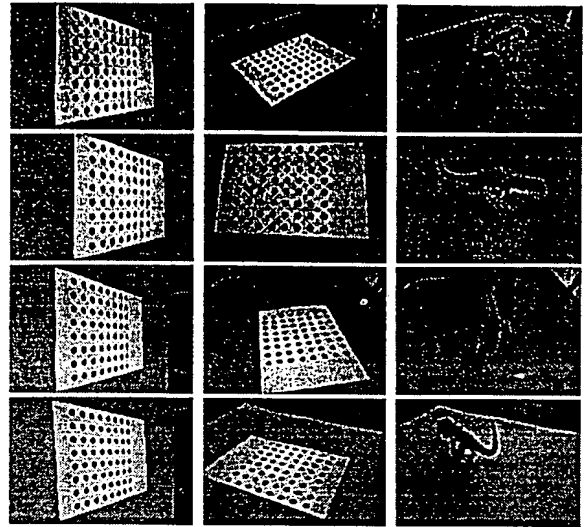


Figure 2: Toy dinosaur images. First column: photographs taken of grid with each camera in portrait orientation (roll angle approximately 90°). Second column: photographs taken with camera in final position in landscape orientation (roll angle approximately 0°). Third column: photographs of toy dinosaur with cameras in final position.

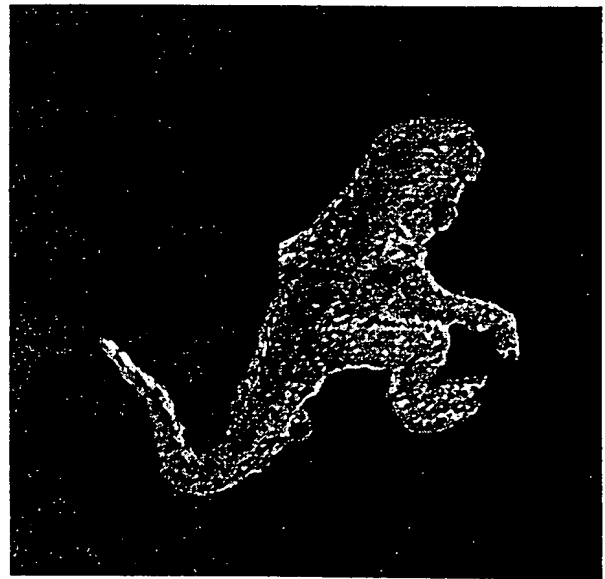


Figure 3: View of the 3D model computed from the toy dinosaur images

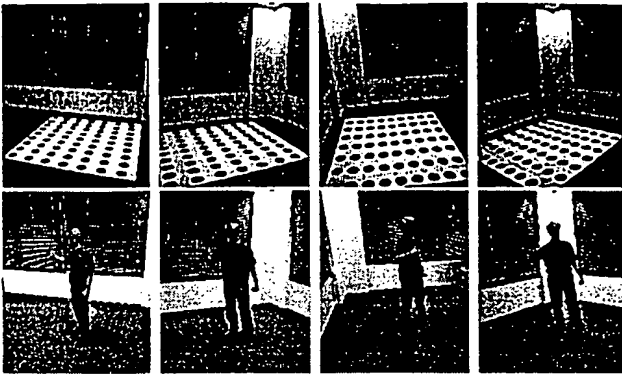


Figure 4: Person images. First row: photographs of the calibration grid with cameras in final position. Second row: photographs of a person.

shown in Figure 5. The accuracy of the camera calibration is demonstrated by a plausible reconstruction of the person.

4 Conclusion

We have shown that accurate camera calibration can be achieved with a simple two views of a plane technique and have demonstrated its practicality by using it to calibrate a multi-camera system for modelling real objects of varying size. Future work will focus on improved techniques for surface generation.

Acknowledgments

The authors would like to thank Ricahrd Taylor, Jane Haslam, Adam Baumberg, Alex Lyons, Simon Rowe and Mike Taylor for their software implementations and Philip McLauchlan for helpful discussions.

References

- [1] O. Faugeras. *Three-Dimensional Computer Vision*. The MIT Press. 1993.
- [2] T. Kanade, P.J. Narayanan, and P.W. Rander. "Virtualized Reality: Constructing virtual worlds from real scenes". *IEEE Multimedia*, 4(1), May 1997.
- [3] A. Katkere, S. Moezzi, D.Y. Kuramura, P. Kelly and R. Jain. "Towards video-based immersive environments". *Multimedia Systems*. 5:69-85, 1997
- [4] R.K. Lenz, and R.Y. Tsai. "Techniques for calibration of the scale factor and image centre for high accuracy 3D machine vision metrology". *Proc. IEEE Int. Conf. Robotics and Automation*, Raleigh, NC, 68-75, March 1987.
- [5] W. Niem and J. Wingbermuhle. "Automatic reconstruction using a mobile monoscopic camera". *Proc. International Conference on Recent Advances in 3D Imaging and Modelling*. Ottawa, Canada, 12-15 May 1997.
- [6] M. Pollefeys, R Koch and L Van Gool. "Self-calibration and metric reconstruction in spite of varying and unknown intrinsic camera parameters". *Proc. International Conference on Computer Vision*, Bombay, India, January 1998.
- [7] R.Y. Tsai. "A versatile camera calibration technique for high accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses". *IEEE Journal of Robotics and Automation*, RA-3(4):323-344, 1987.
- [8] C. Wiles. "Calibrating and 3D modelling with a multi-camera system". Canon Technical Report CRE-TR-98-043, 14 December 1998.
- [9] Geometrix Inc. <http://www.geometrixinc.com/>
- [10] Oxford Metrics. <http://www.oxfordmetrics.co.uk/>
- [11] Vanguard. <http://www.robots.ox.ac.uk/~vanguard/>



Figure 5: View of the 3D model computed from person images

CLAIMS

1. Apparatus for processing image data and sound data, comprising:

5 image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

sound processing means for processing sound data
10 to determine the direction of arrival of the sound;

speaker identification means for determining which of the people is speaking based on the result of the processing performed by the image processing means and the result of the processing performed by the sound
15 processing means; and

voice recognition processing means for processing the received sound data to generate text data therefrom in dependence upon the result of the processing performed by the speaker identification means.

20

2. Apparatus according to claim 1, wherein the voice recognition processing means includes storage means for storing respective voice recognition parameters for each of the people, and means for selecting the voice
25 recognition parameters to be used to process the sound data in dependence upon the person determined to be

speaking by the speaker identification means.

3. Apparatus according to claim 1 or claim 2, wherein
the image processing means is arranged to track each
5 person by processing the image data using camera
calibration data defining the position and orientation
of each camera from which image data is processed.

4. Apparatus according to any preceding claim, wherein
10 the image processing means is arranged to track each
person by tracking each person's head.

5. Apparatus according to any preceding claim, wherein
the image processing means is arranged to process the
15 image data to determine where at least each person who
is speaking is looking.

6. Apparatus according to any preceding claim, wherein
the speaker identification means is arranged to identify
20 a person who is speaking in a given frame of the received
image data using the results of the processing performed
by the image processing means and the sound processing
means for at least one other frame if the speaker cannot
be identified using the results of the processing
25 performed by the image processing means and the sound
processing means for the given frame.

7. Apparatus according to any preceding claim, further comprising a database for storing at least some of the received image data, the sound data, the text data produced by the voice recognition processing means and
5 viewing data defining where at least each person who is speaking is looking, the database being arranged to store the data such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

10

8. Apparatus according to claim 7, further comprising means for compressing the image data and the sound data for storage in the database.

15 9. Apparatus according to claim 8, wherein the means for compressing the image data and the sound data comprises means for encoding the image data and the sound data as MPEG data.

20 10. Apparatus according to any of claims 7 to 9, further comprising means for generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and wherein the database is
25 arranged to store the data so that it is associated with the corresponding image data, sound data, text data and

viewing data.

11. Apparatus according to claim 10, wherein the predetermined period comprises a period during which the given person was talking.

12. Apparatus for processing image data and sound data, comprising:

image processing means for processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

sound processing means for processing sound data to determine the direction of arrival of the sound;

speaker identification means for determining which of the people is speaking based on the result of the processing performed by the image processing means and the result of the processing performed by the sound processing means.

20

13. Apparatus according to claim 12, wherein the image processing means is arranged to track each person by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

25

14. Apparatus according to claim 12 or claim 13, wherein the image processing means is arranged to track each person by tracking each person's head.

5 15. Apparatus according to any of claims 12 to 14, wherein the image processing means is arranged to process the image data to determine where at least each person who is speaking is looking.

10 16. Apparatus according to any of claims 12 to 15, wherein the speaker identification means is arranged to identify a person who is speaking in a given frame of the received image data using the results of the processing performed by the image processing means and the sound
15 processing means for at least one other frame if the speaker cannot be identified using the results of the processing performed by the image processing means and the sound processing means for the given frame.

20 17. A method of processing image data and sound data, comprising:

an image processing step of processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three
25 dimensions;

a sound processing step of processing sound data

to determine the direction of arrival of the sound;

a speaker identification step of determining which of the people is speaking based on the result of the processing performed in the image processing step and the
5 result of the processing performed in the sound processing step; and

a voice recognition processing step of processing the received sound data to generate text data therefrom in dependence upon the result of the processing performed
10 in the speaker identification step.

18. A method according to claim 17, wherein, the voice recognition processing step includes selecting, from stored respective voice recognition parameters for each
15 of the people, the voice recognition parameters to be used to process the sound data in dependence upon the person determined to be speaking in the speaker identification step.

20 19. A method according to claim 17 or claim 18, wherein, in the image processing step, each person is tracked by processing the image data using camera calibration data defining the position and orientation of each camera from which image data is processed.

25

20. A method according to any of claims 17 to 19,

wherein, in the image processing step, each person is tracked by tracking the person's head.

21. A method according to any of claims 17 to 20,
5 wherein, in the image processing step, the image data is processed to determine where at least each person who is speaking is looking.

22. A method according to any of claims 17 to 21,
10 wherein, in the speaker identification step, a person who is speaking in a given frame of the received image data is identified using the results of the processing performed in the image processing step and the sound processing step for at least one other frame if the
15 speaker cannot be identified using the results of the processing performed in the image processing step and the sound processing step for the given frame.

23. A method according to any of claims 17 to 22,
20 further comprising the step of generating a signal conveying the text data generated in the voice recognition processing step.

24. A method according to any of claims 17 to 23,
25 further comprising the step of storing in a database at least some of the received image data, the sound data,

the text data produced in the voice recognition processing step and viewing data defining where at least each person who is speaking is looking, the data being stored in the database such that corresponding text data and viewing data are associated with each other and with the corresponding image data and sound data.

25. A method according to claim 24, wherein the image data and the sound data are stored in the database in compressed form.

26. A method according to claim 25, wherein the image data and the sound data are stored as MPEG data.

27. A method according to any of claims 24 to 26, further comprising the steps of generating data defining, for a predetermined period, the proportion of time spent by a given person looking at each of the other people during the predetermined period, and storing the data in the database so that it is associated with the corresponding image data, sound data, text data and viewing data.

28. A method according to claim 27, wherein the predetermined period comprises a period during which the given person was talking.

29. A method according to any of claims 24 to 28, further comprising the step of generating a signal conveying the database with data therein.

5 30. A method according to claim 29, further comprising the step of recording the signal either directly or indirectly to generate a recording thereof.

31. A method of processing image data and sound data,
10 comprising:

an image processing step of processing image data recorded by at least one camera showing the movements of a plurality of people to track each person in three dimensions;

15 a sound processing step of processing sound data to determine the direction of arrival of the sound; and

a speaker identification step of determining which of the people is speaking based on the result of the processing performed in the image processing step and the
20 result of the processing performed in the sound processing step.

32. A method according to claim 31, wherein, in the image processing step, each person is tracked by
25 processing the image data using camera calibration data defining the position and orientation of each camera from

which image data is processed.

33. A method according to claim 31 or claim 32, wherein,
in the image processing step, each person is tracked by
5 tracking the person's head.

34. A method according to any of claims 31 to 33,
wherein, in the image processing step, the image data is
processed to determine where at least each person who is
10 speaking is looking.

35. A method according to any of claims 31 to 34,
wherein, in the speaker identification step, a person who
is speaking in a given frame of the received image data
15 is identified using the results of the processing
performed in the image processing step and the sound
processing step for at least one other frame if the
speaker cannot be identified using the results of the
processing performed in the image processing step and the
20 sound processing step for the given frame.

36. A method according to any of claims 31 to 35,
further comprising the step of generating a signal
conveying the identity of the speaker identified in the
25 speaker identification step.

37. A storage device storing instructions for causing a programmable processing apparatus to become configured as an apparatus as set out in any of claims 1 to 16.

5 38. A storage device storing instructions for causing a programmable processing apparatus to become operable to perform a method as set out in any of claims 17 to 36.

39. A signal conveying instructions for causing a
10 programmable processing apparatus to become configured as an apparatus as set out in any of claims 1 to 16.

40. A signal conveying instructions for causing a
15 programmable processing apparatus to become operable to perform a method as set out in any of claims 17 to 36.

ABSTRACTIMAGE AND SOUND PROCESSING APPARATUS

Image data from a plurality of cameras 2-1, 2-2, 2-3
5 showing the movements of a number of people, for example
in a meeting, and sound data from a directional
microphone array 4 is processed by a computer processing
apparatus 24 to archive the data in a meeting archive
database 60. The image data is processed to determine
10 the three-dimensional position and orientation of each
person's head and to determine at whom each person is
looking. The sound data is processed to determine the
direction from which the sound came. Processing is
carried out to determine who is speaking by determining
15 which person has his head in a position corresponding to
the direction from which the sound came. Having
determined which person is speaking, the personal speech
recognition parameters for that person are selected and
used to convert the sound data to text data. Image data
20 to be archived is chosen by selecting the camera which
best shows the speaking participant and the participant
to whom he is speaking. Image data, sound data, text
data and data defining at whom each person is looking is
stored in the meeting archive database 60.

25

(FIGURE 2)

THIS PAGE BLANK (USPTO)

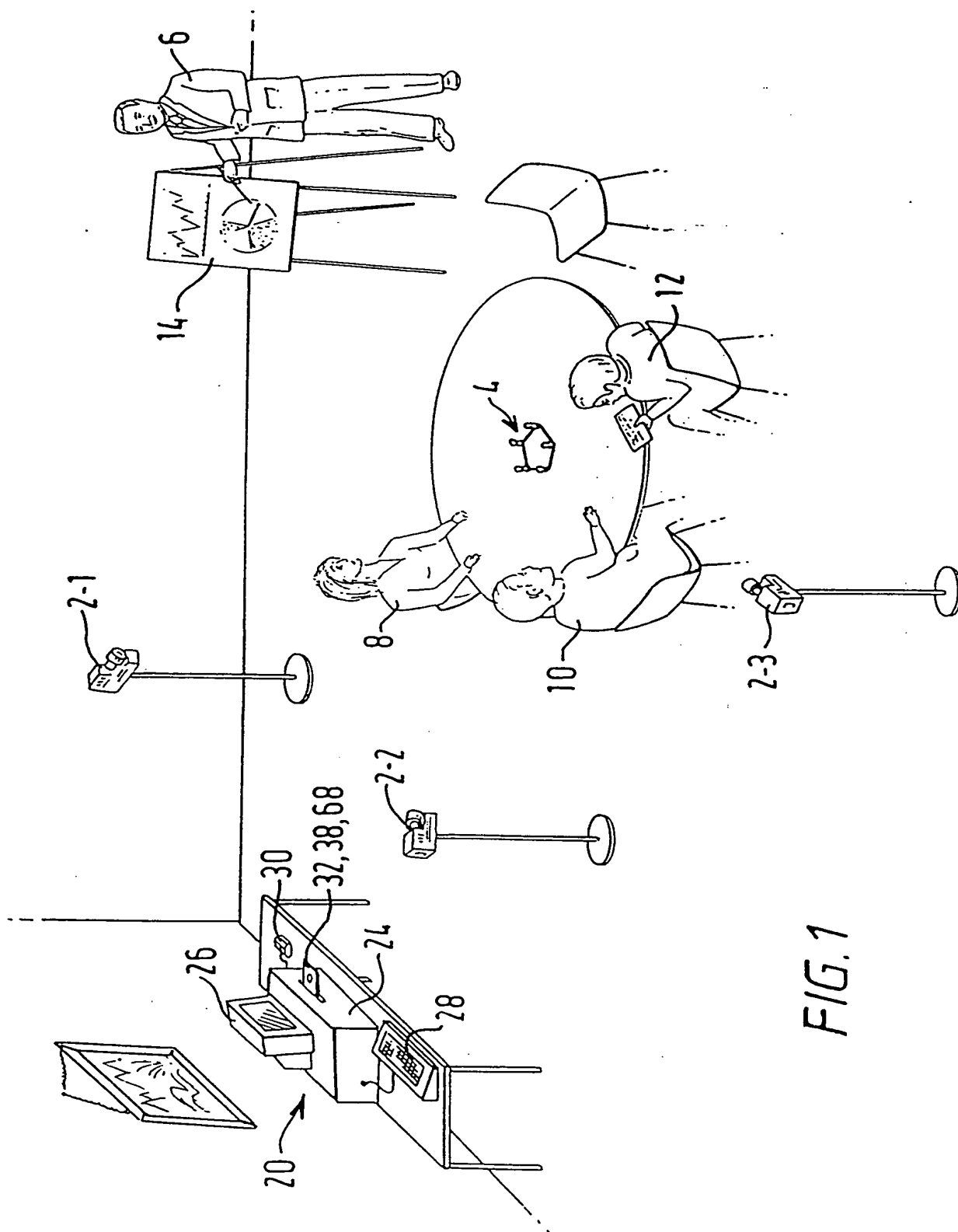


FIG. 1

THIS PAGE BLANK (USPTO)

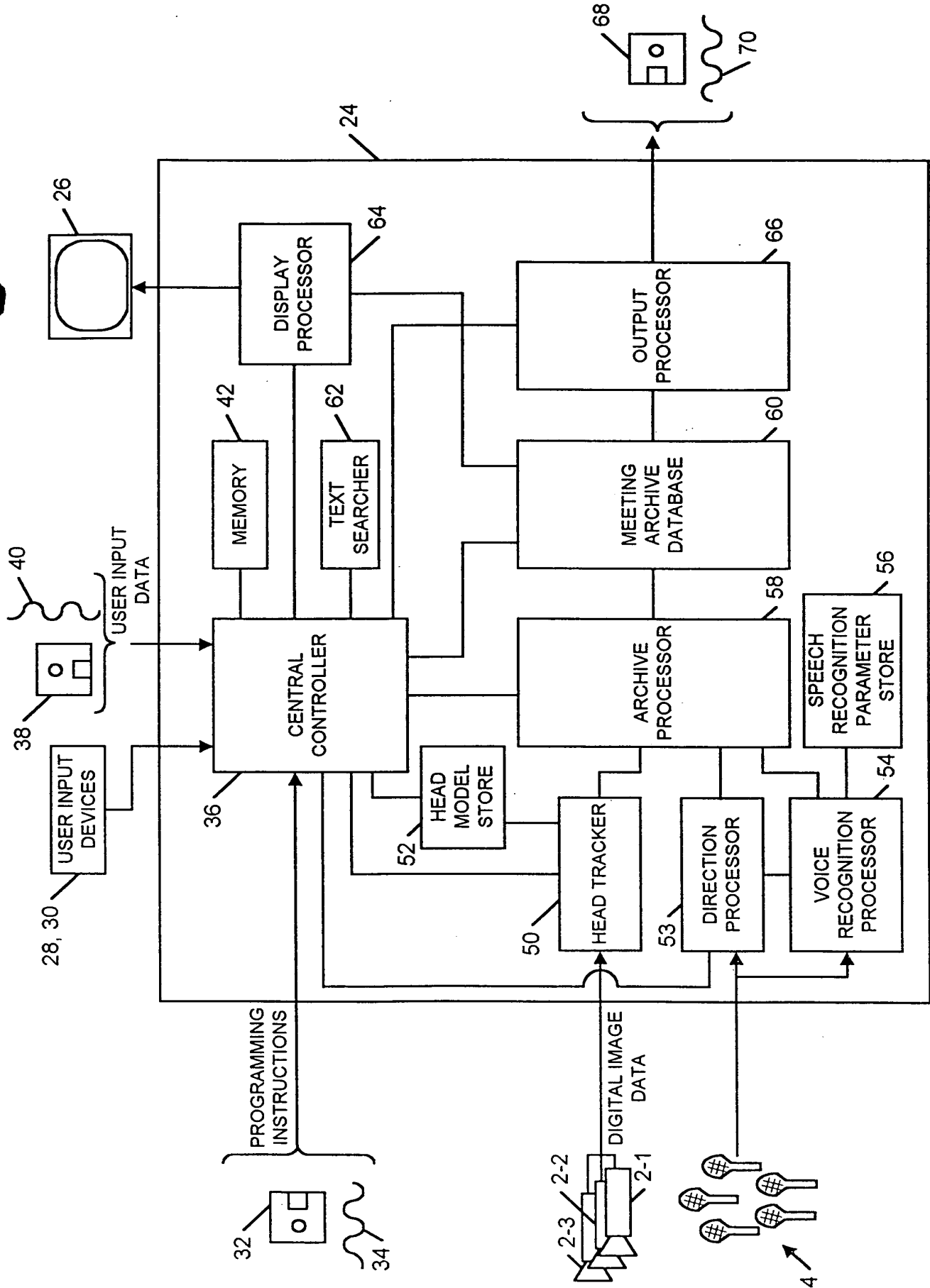
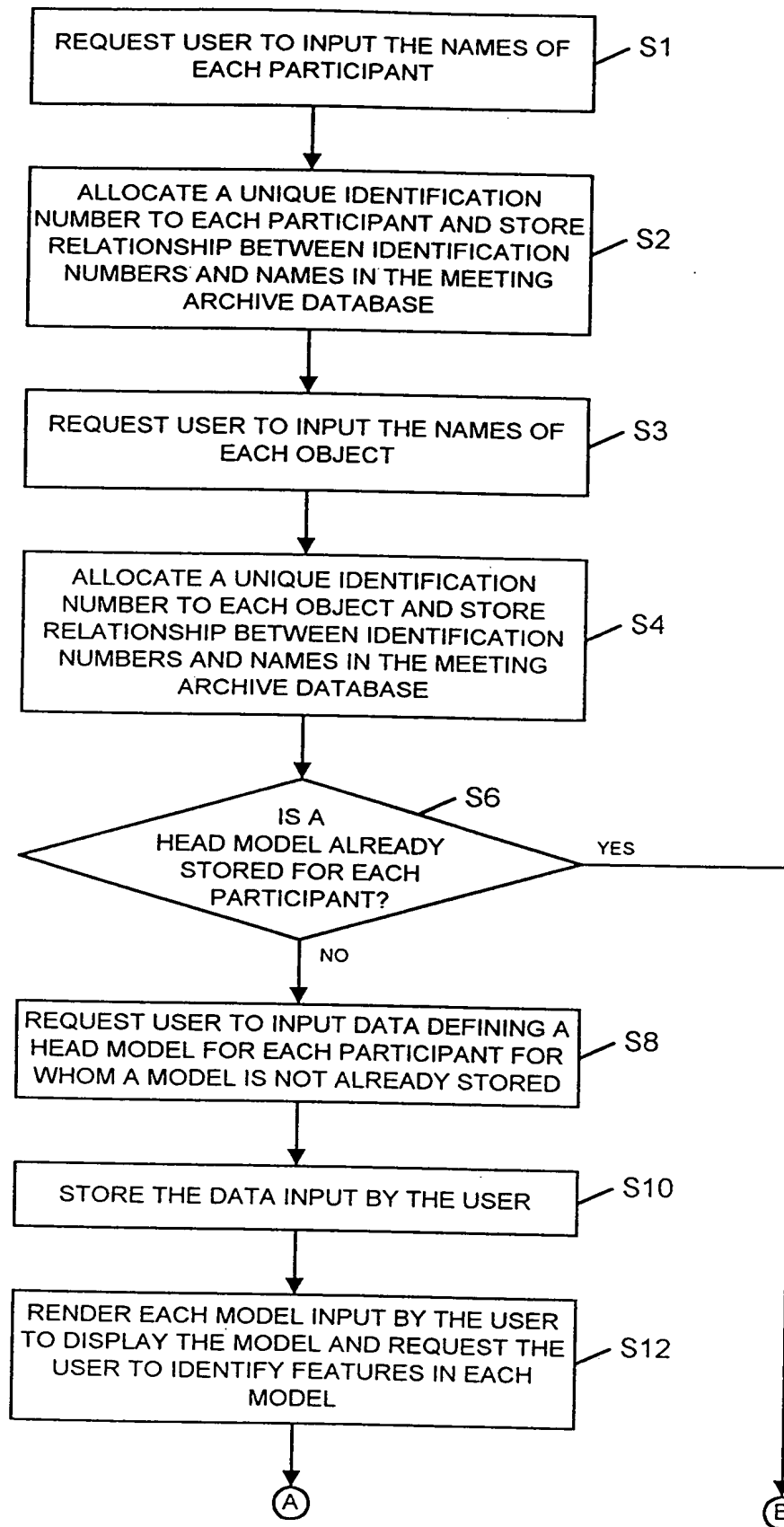


FIG. 2

THIS PAGE BLANK (USPTO)



THIS PAGE BLANK (USPTO)

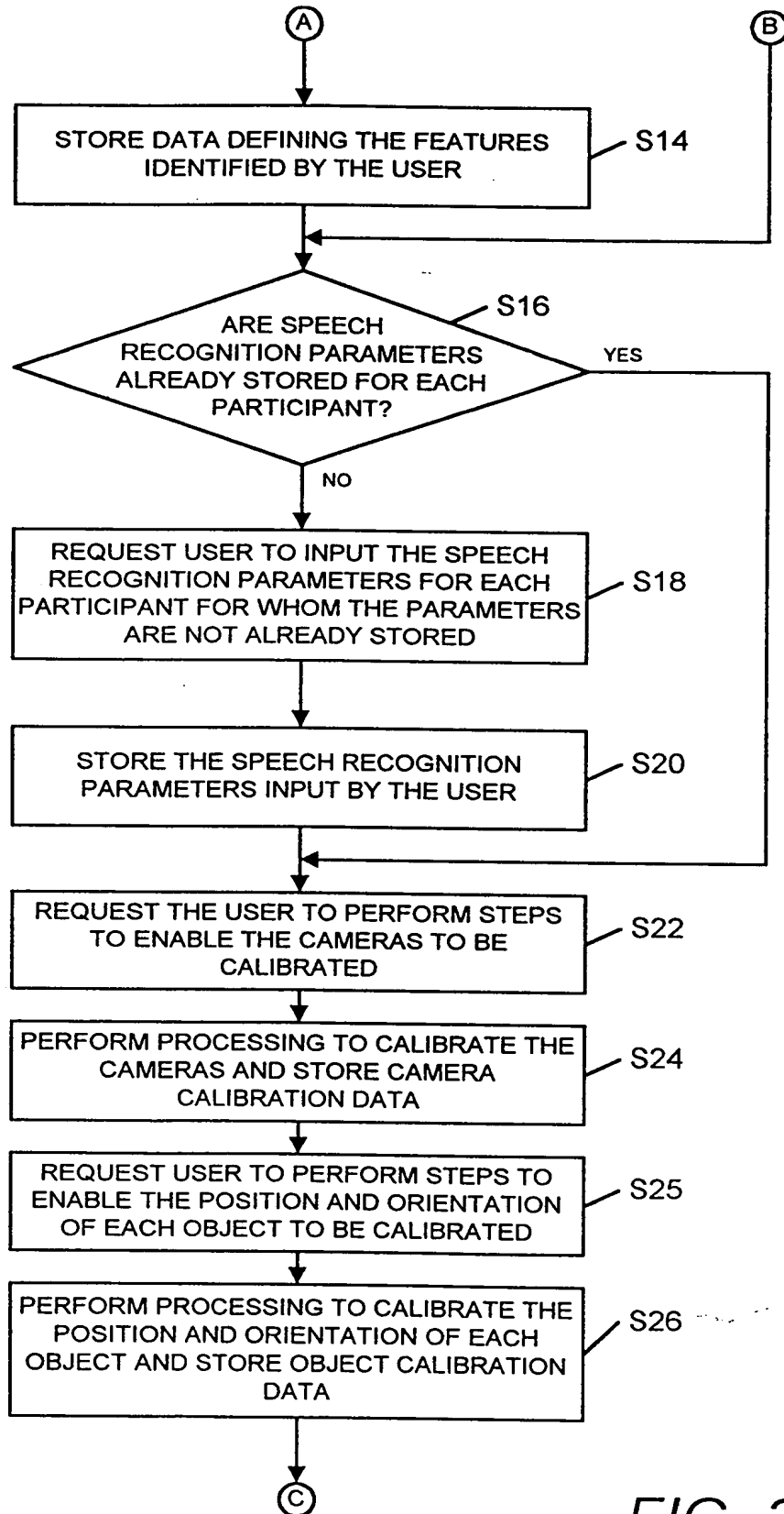


FIG. 3 (cont)

THIS PAGE BLANK (USPTO)

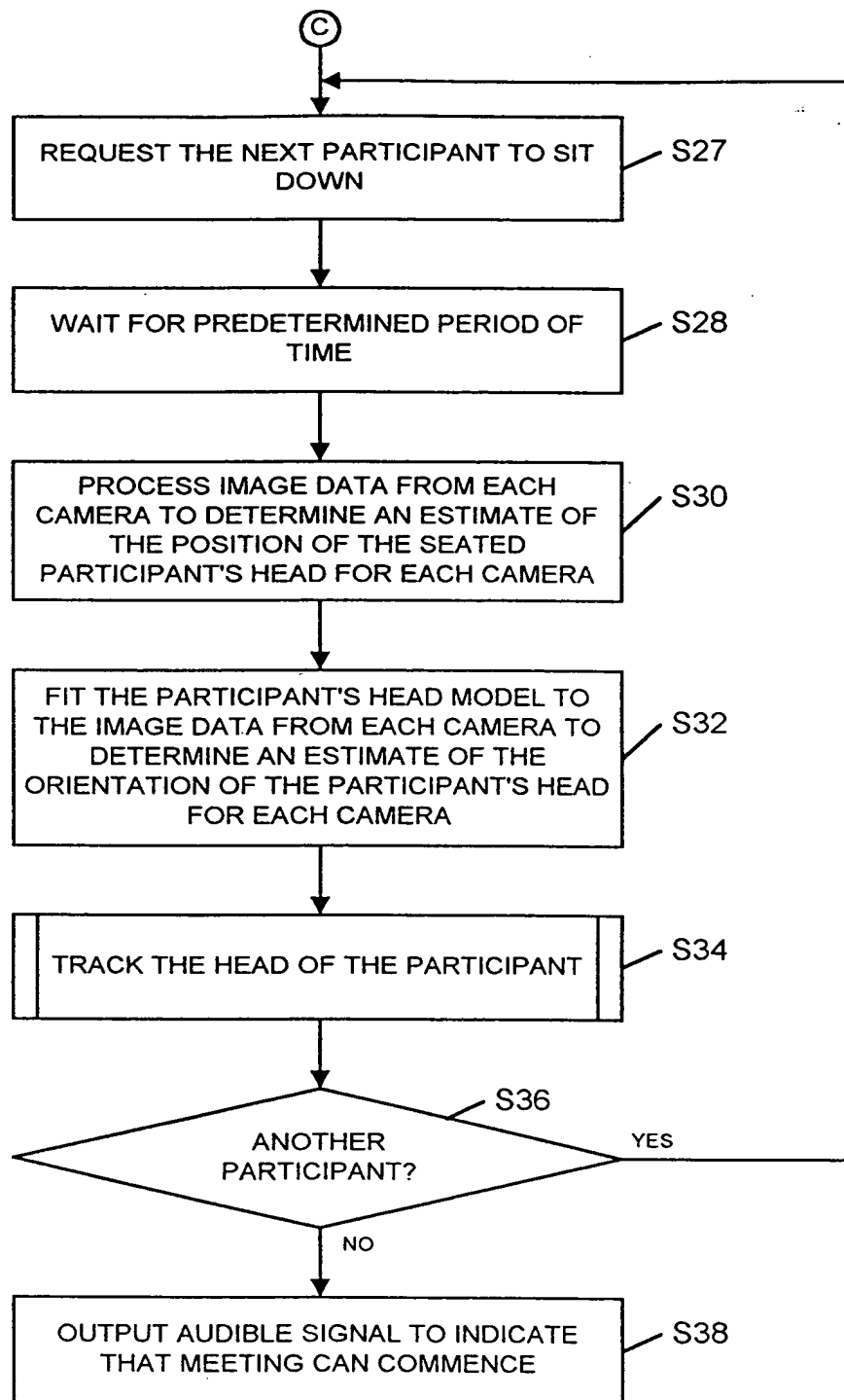


FIG. 3 (cont)

THIS PAGE BLANK (USPTO)

6/24

NUMBER	NAME
1	MR. A
2	MISS. B
3	MR. C
4	MISS. D
5	FLIP CHART

80

FIG. 4

THIS PAGE BLANK (USPTO)

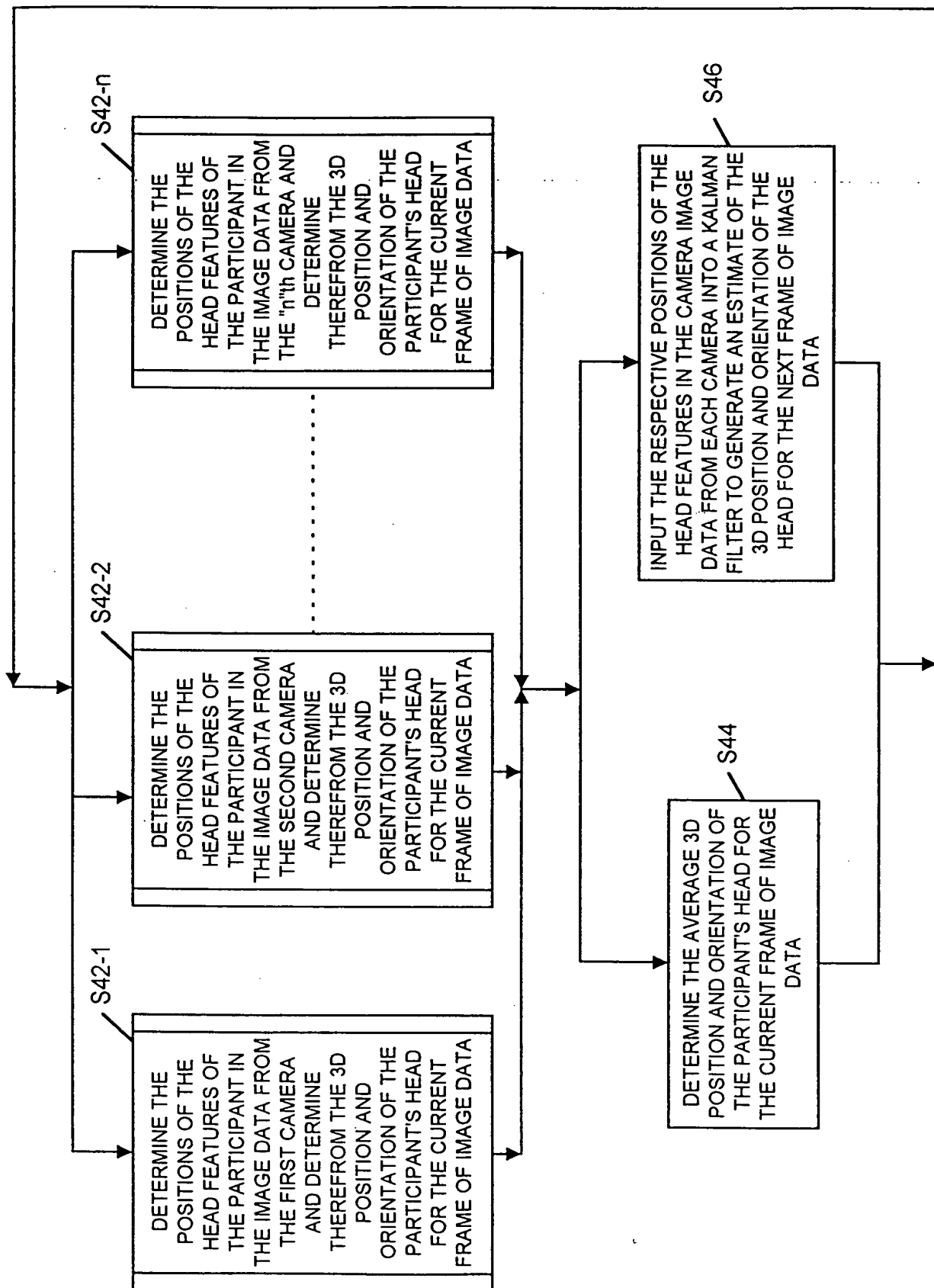


FIG. 5

THIS PAGE BLANK (USPTO)

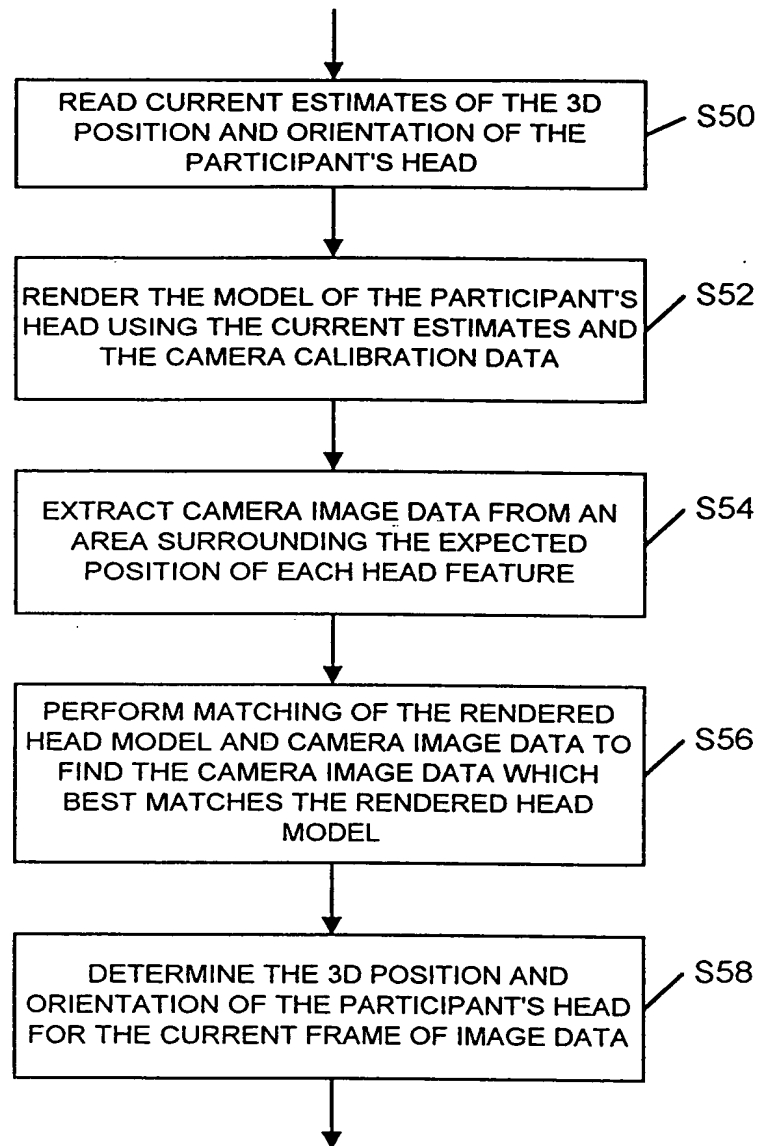


FIG. 6

THIS PAGE BLANK (USPTO)

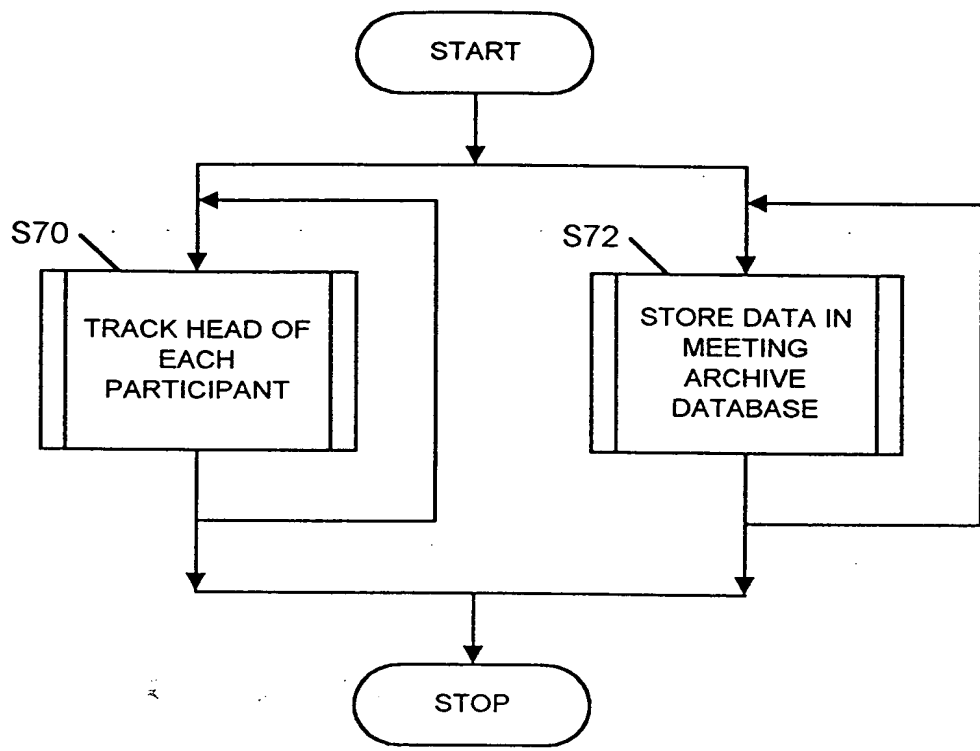


FIG. 7

THIS PAGE BLANK (USPTO)

10/24

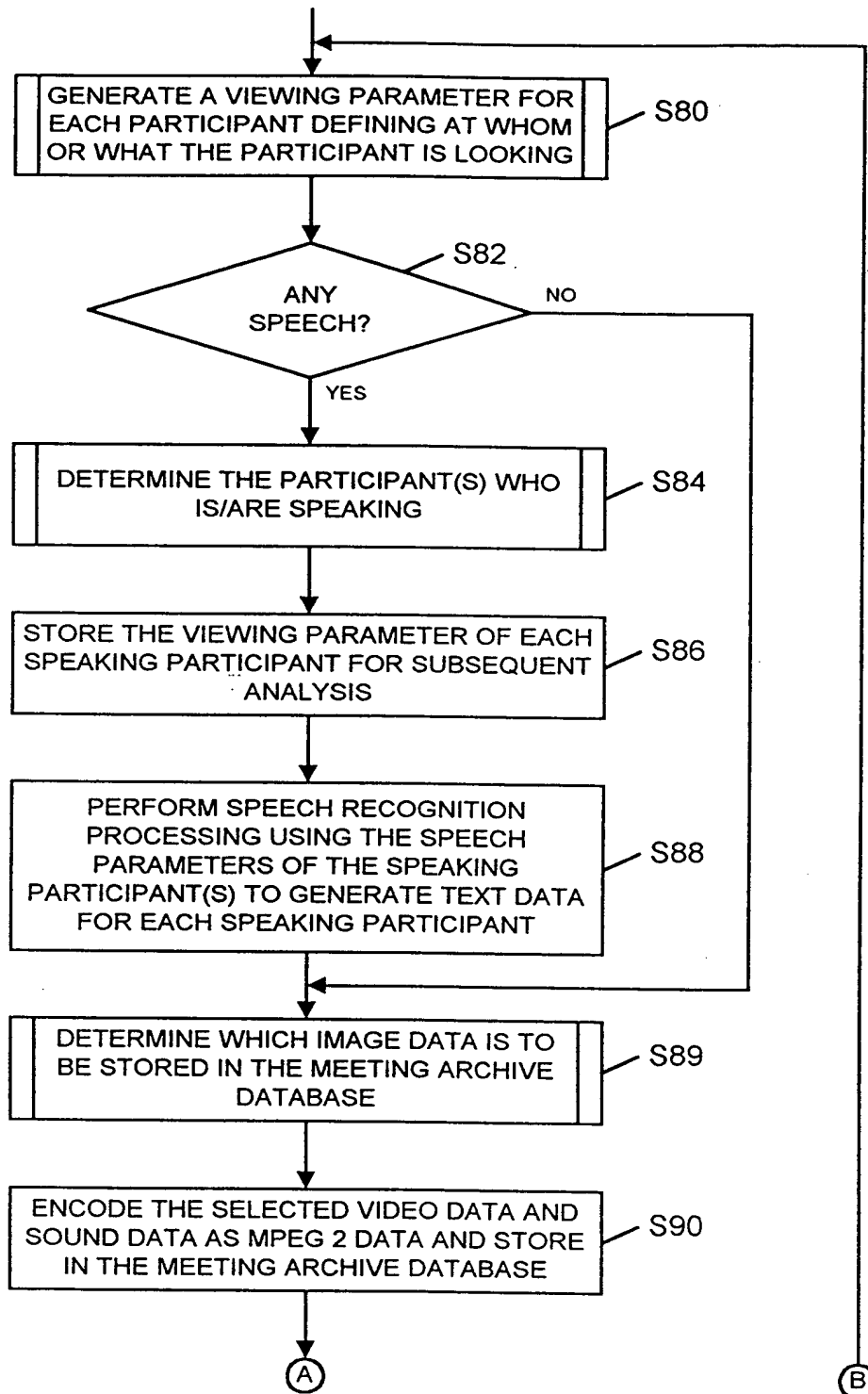


FIG. 8

THIS PAGE BLANK (USPTO)

11/24

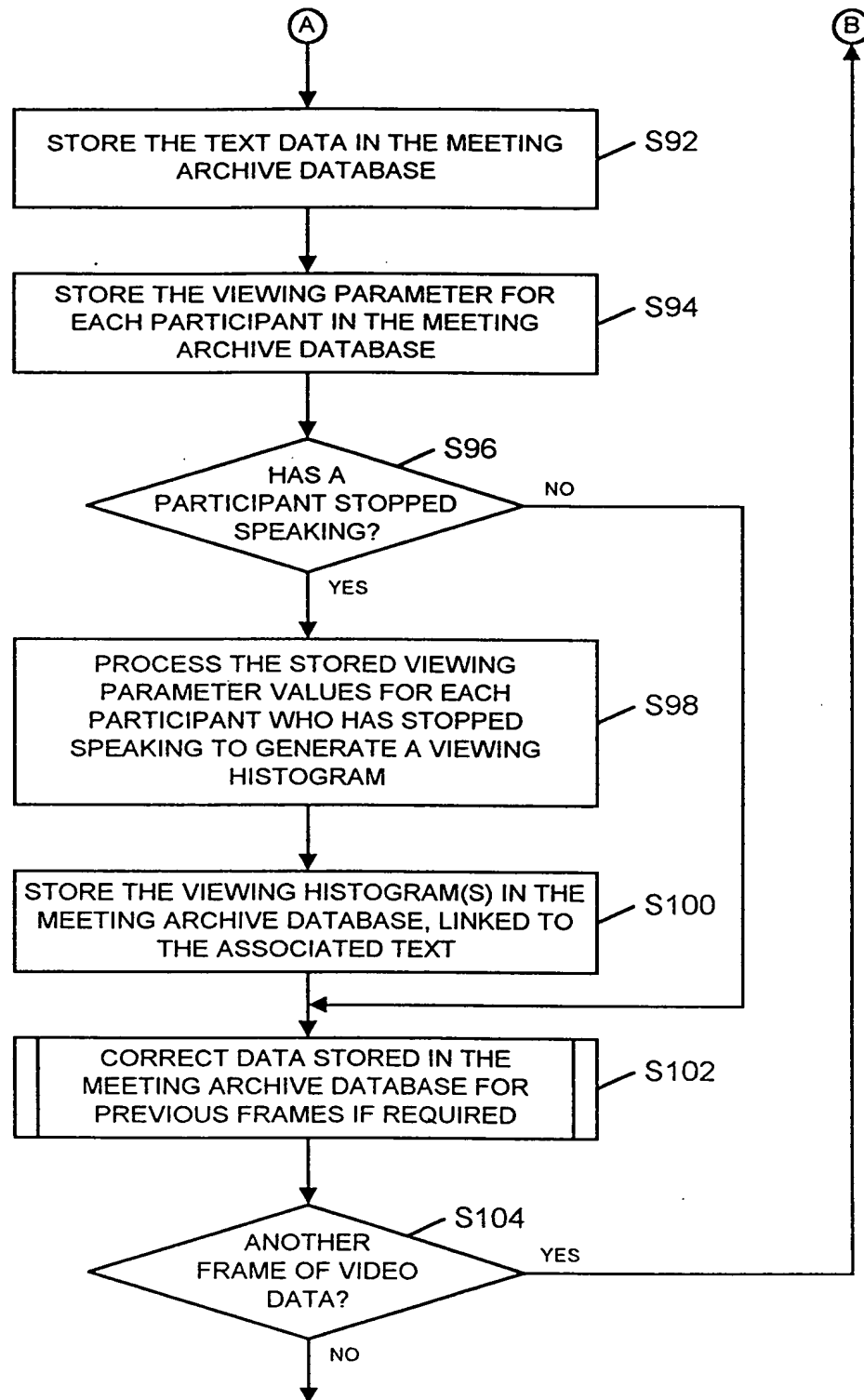
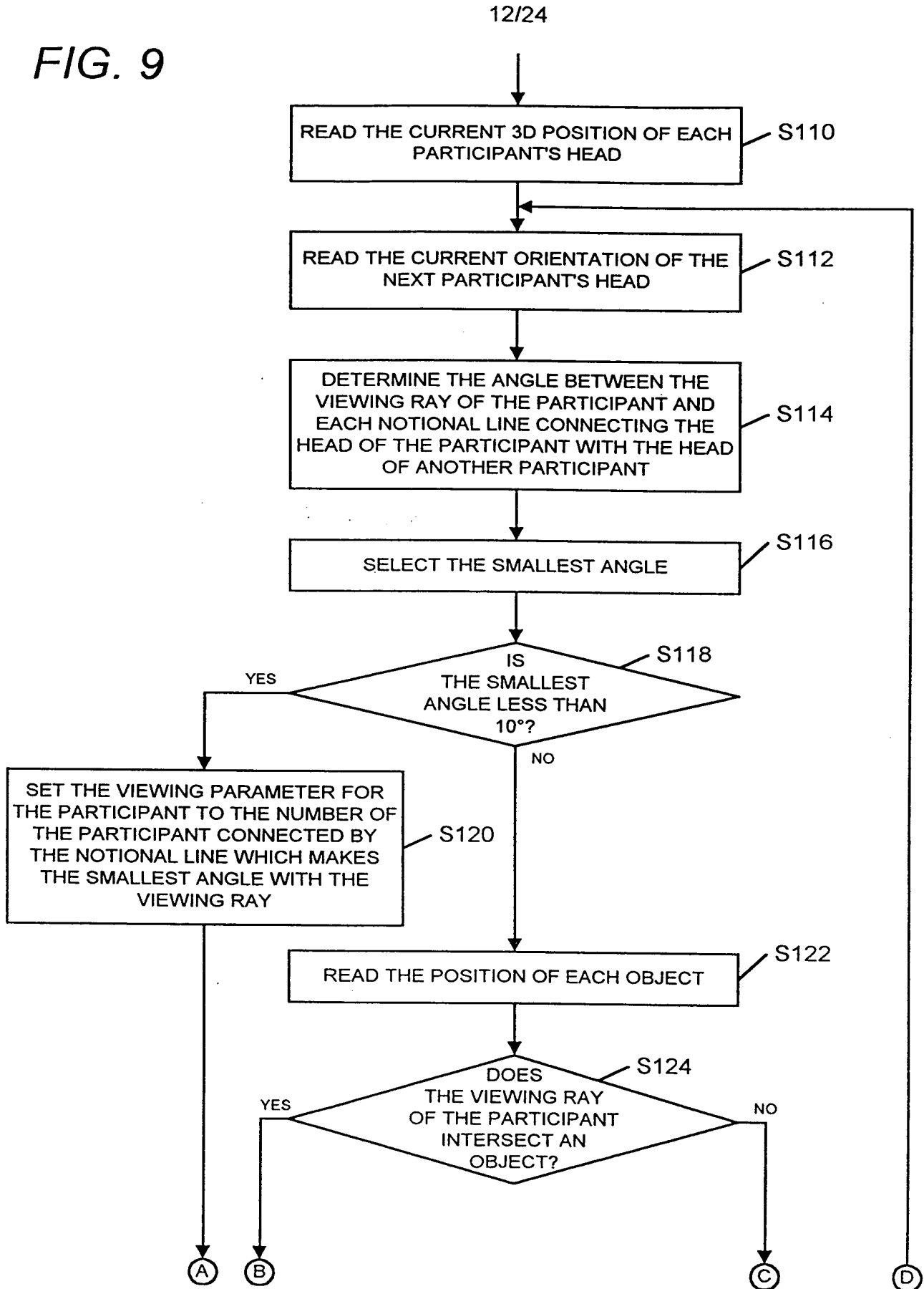


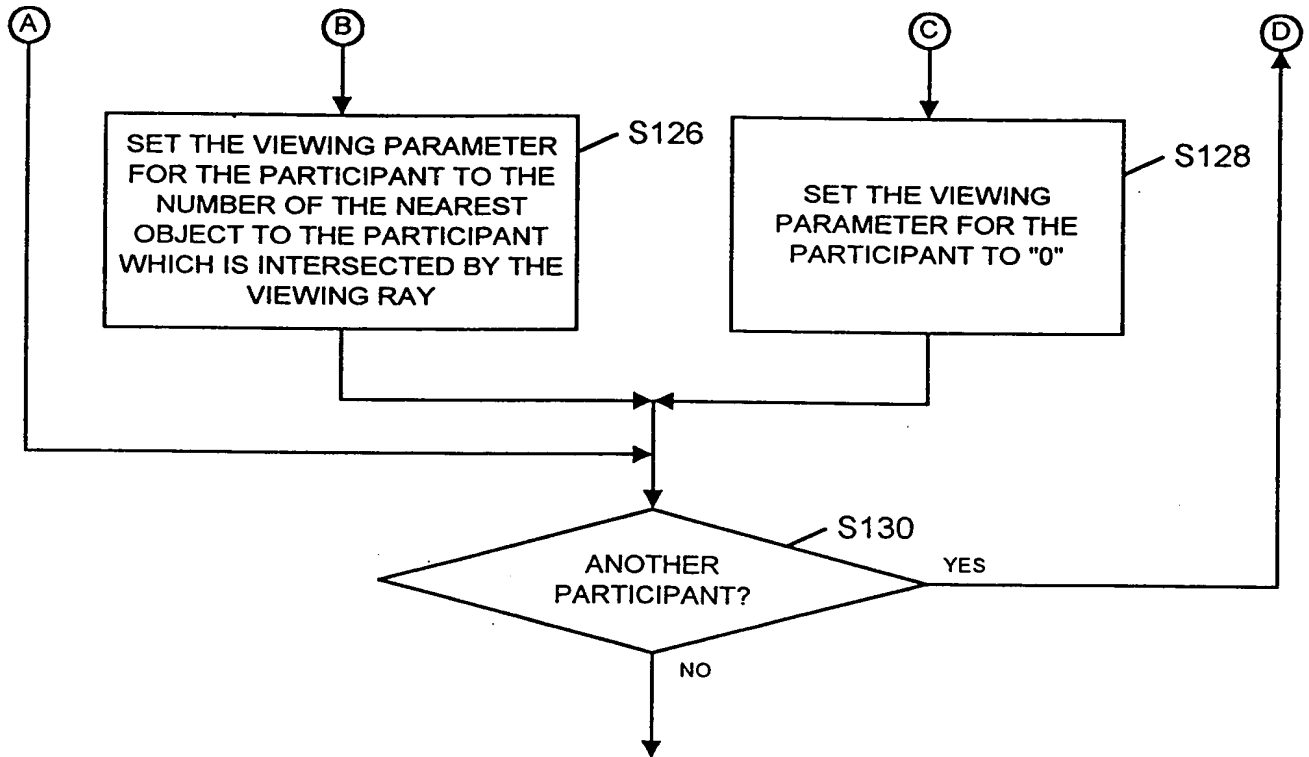
FIG. 8 (cont)

THIS PAGE BLANK (USPTO)

FIG. 9

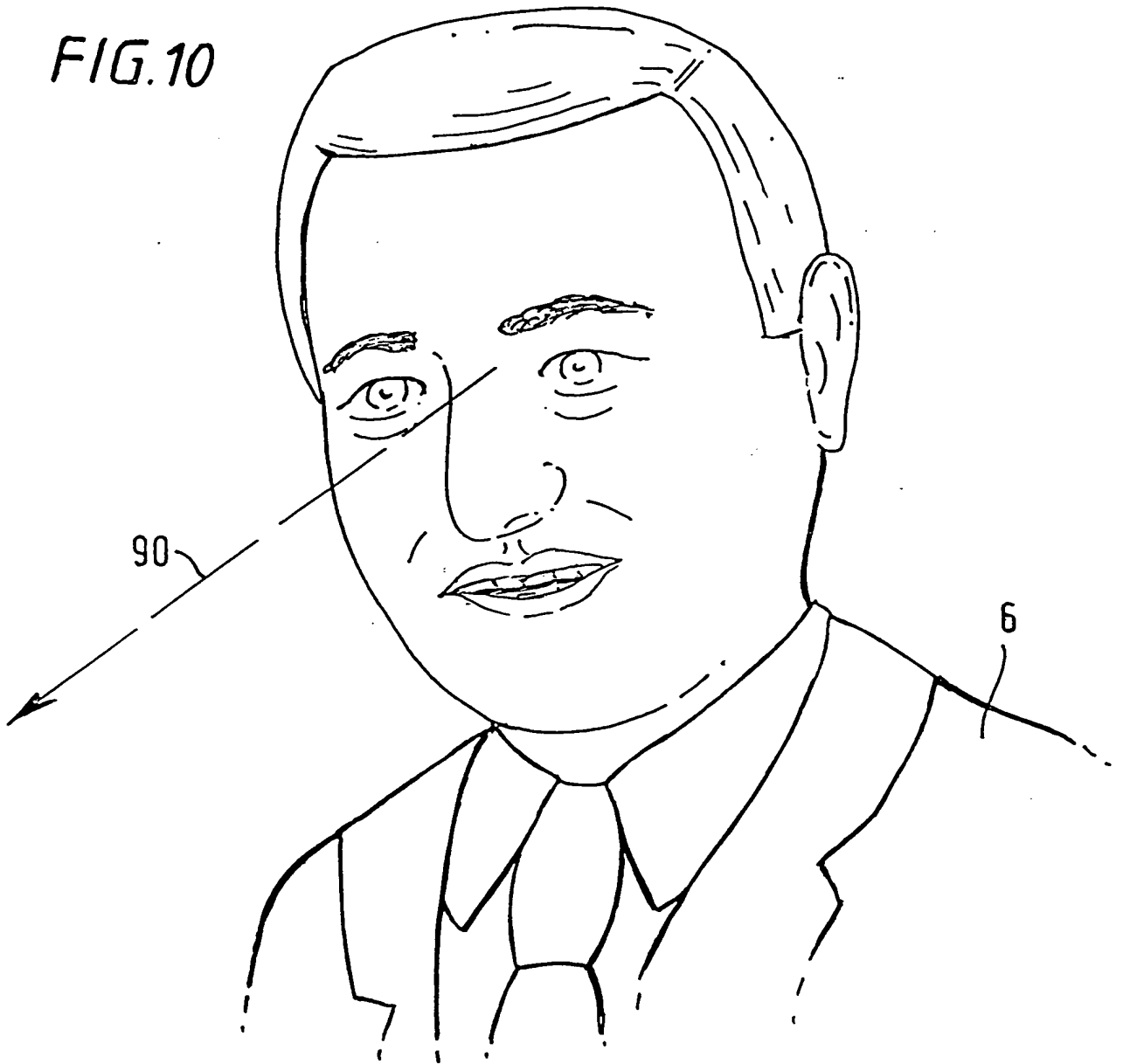


THIS PAGE BLANK (USPTO)

*FIG. 9 (cont)*

THIS PAGE BLANK (USPTO)

FIG. 10



THIS PAGE BLANK (USPTO)

THIS PAGE BLANK (USPTO)

16/24

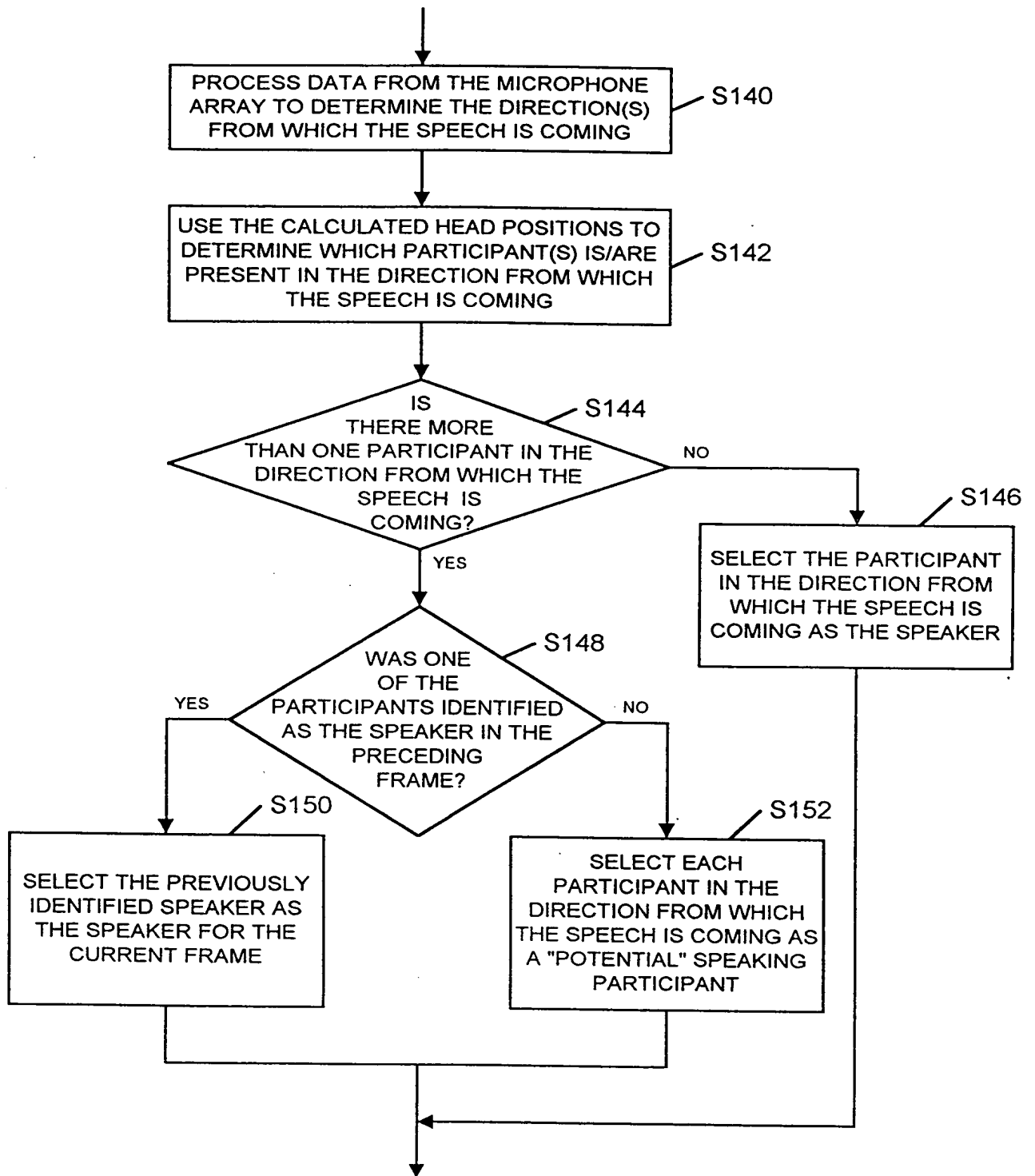


FIG. 12

THIS PAGE BLANK (USPTO)

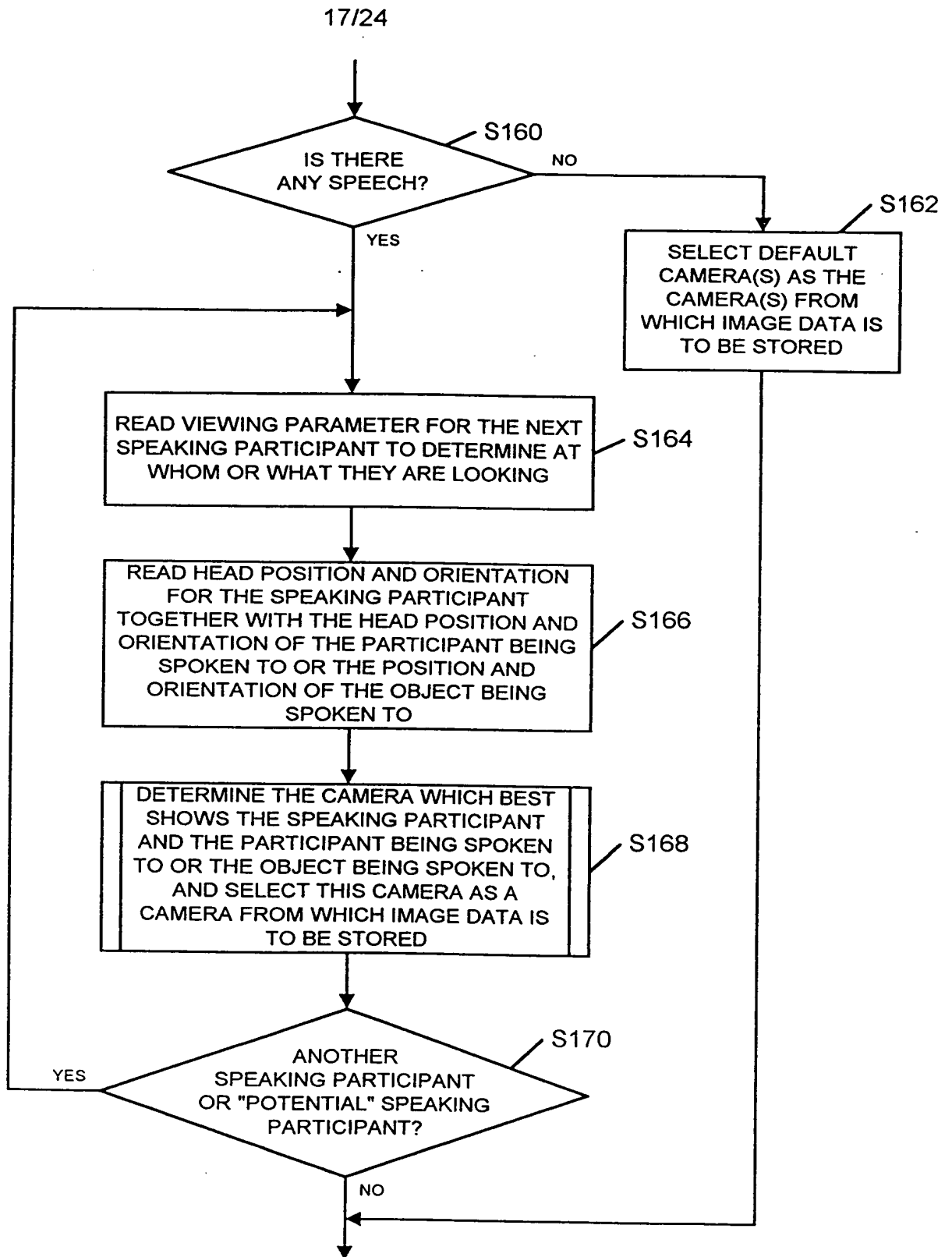


FIG. 13

THIS PAGE BLANK (USPTO)

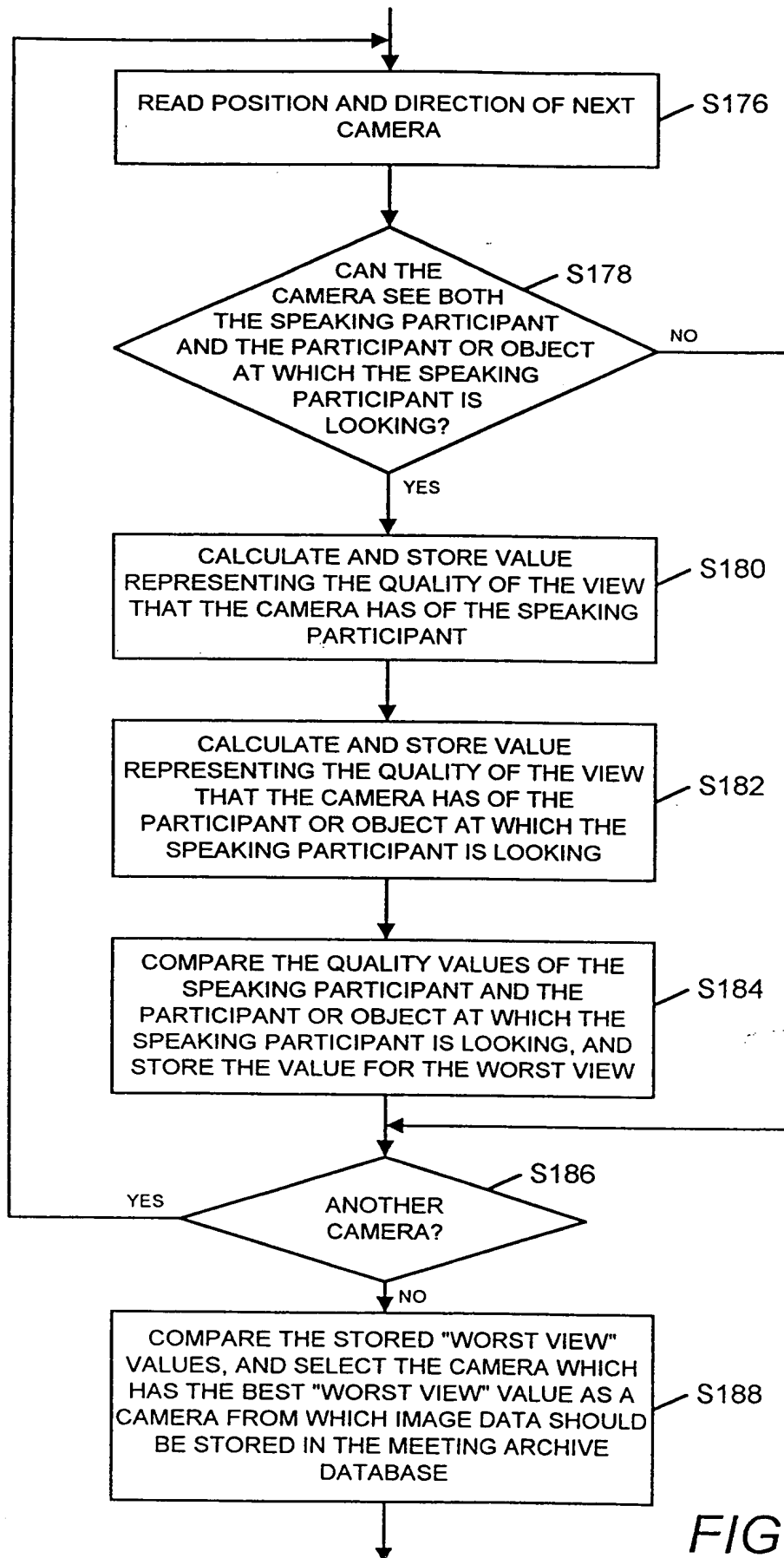
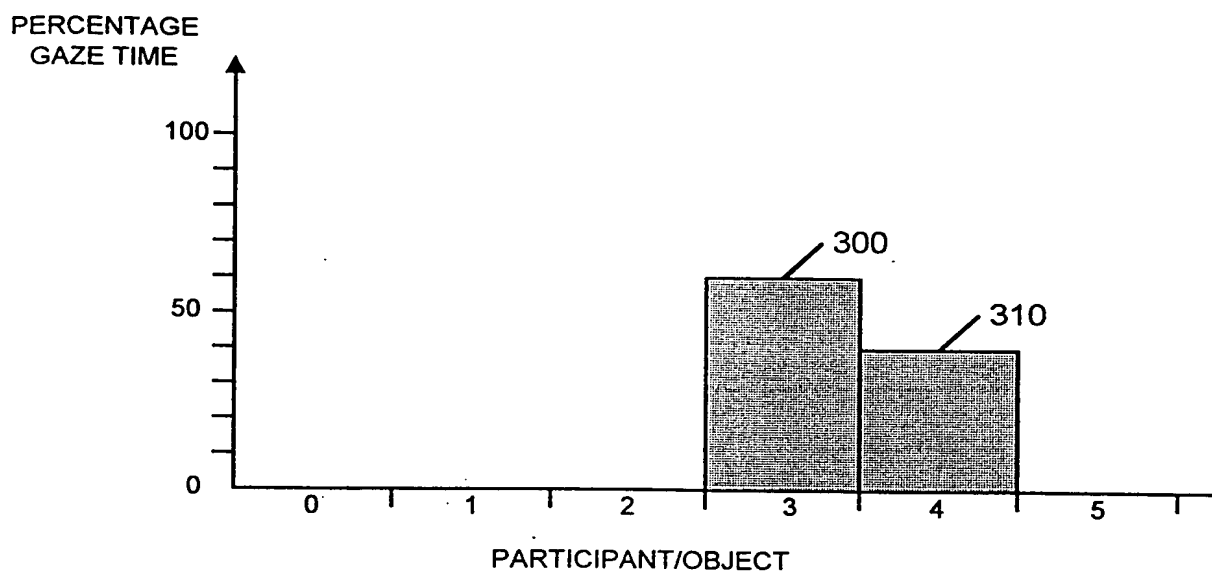
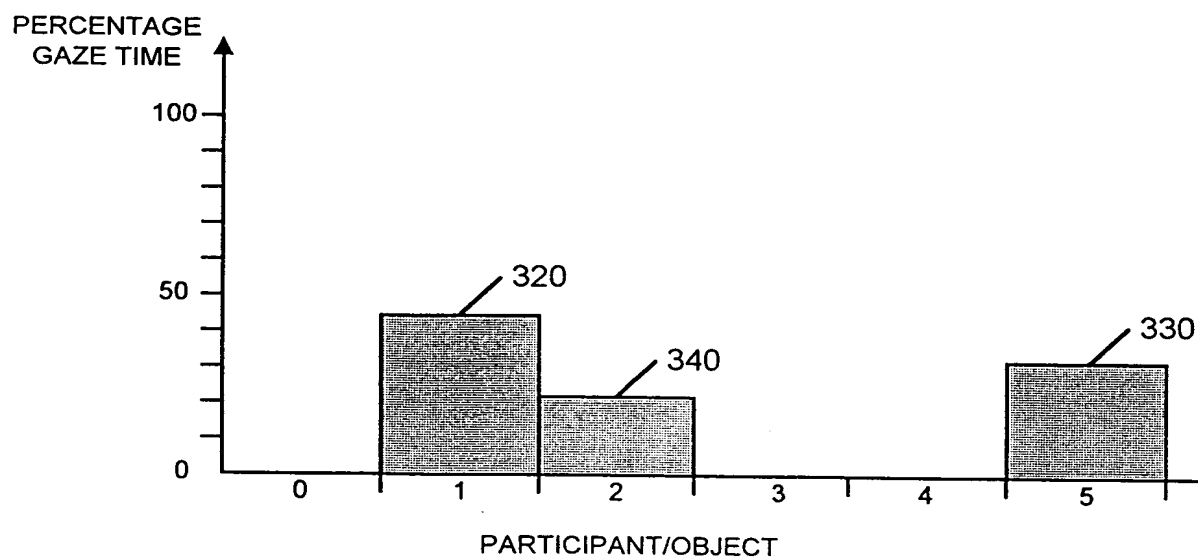


FIG. 14

THIS PAGE BLANK (USPTO)

THIS PAGE BLANK (USPTO)

*FIG. 16A**FIG. 16B*

THIS PAGE BLANK (USPTO)

21/24

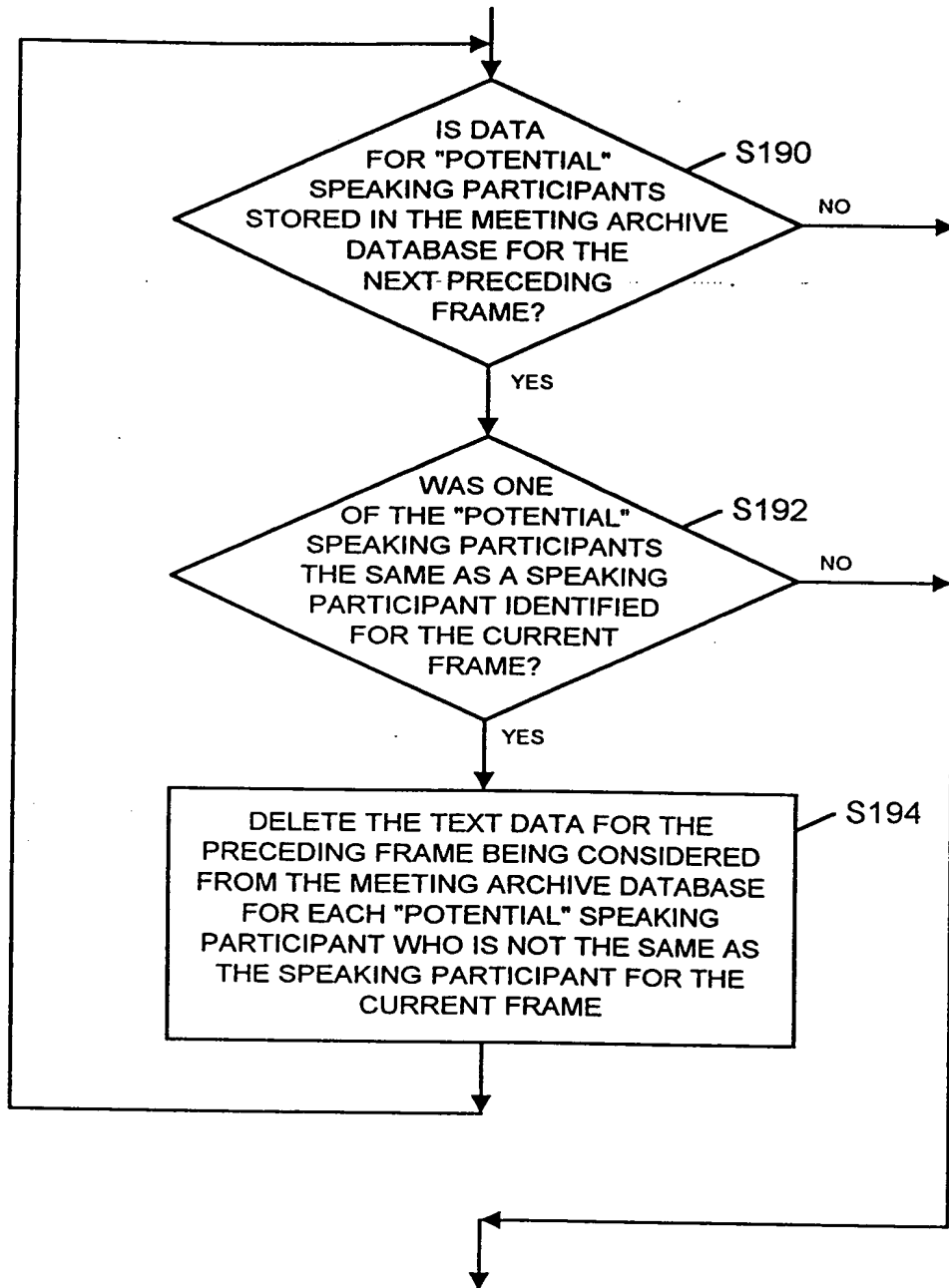


FIG. 17

THIS PAGE BLANK (USPTO)

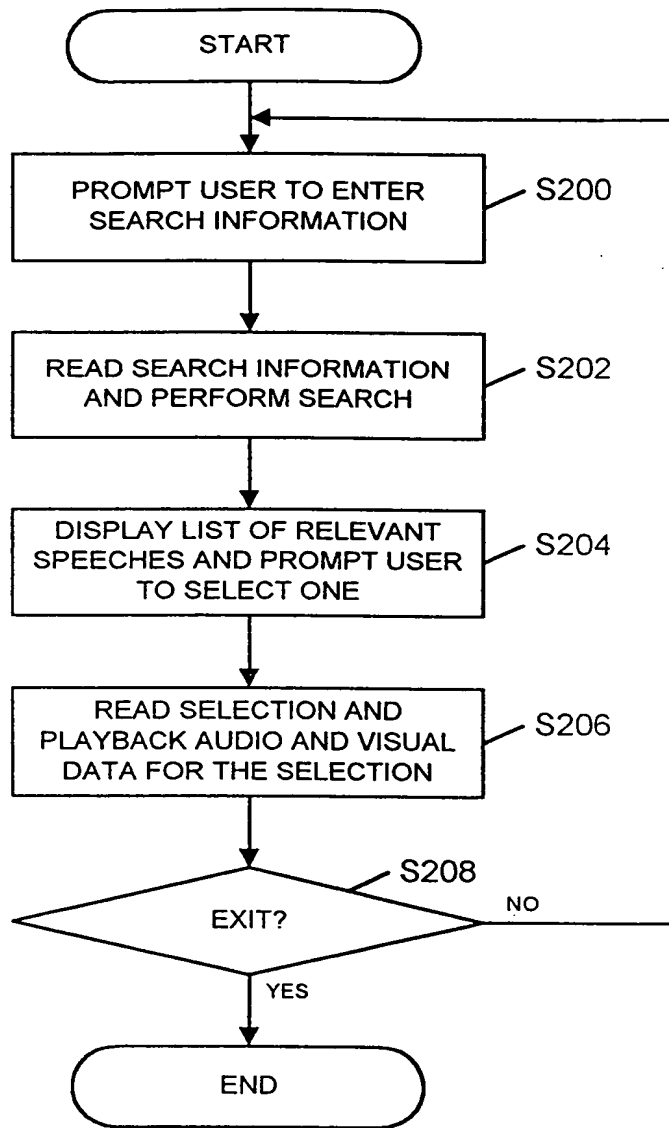


FIG. 18

THIS PAGE BLANK (USPTO)

Please enter search parameters

400 talking about 410 to 420

Time limits:

Before 430

After 440

Between 450 and 460

470

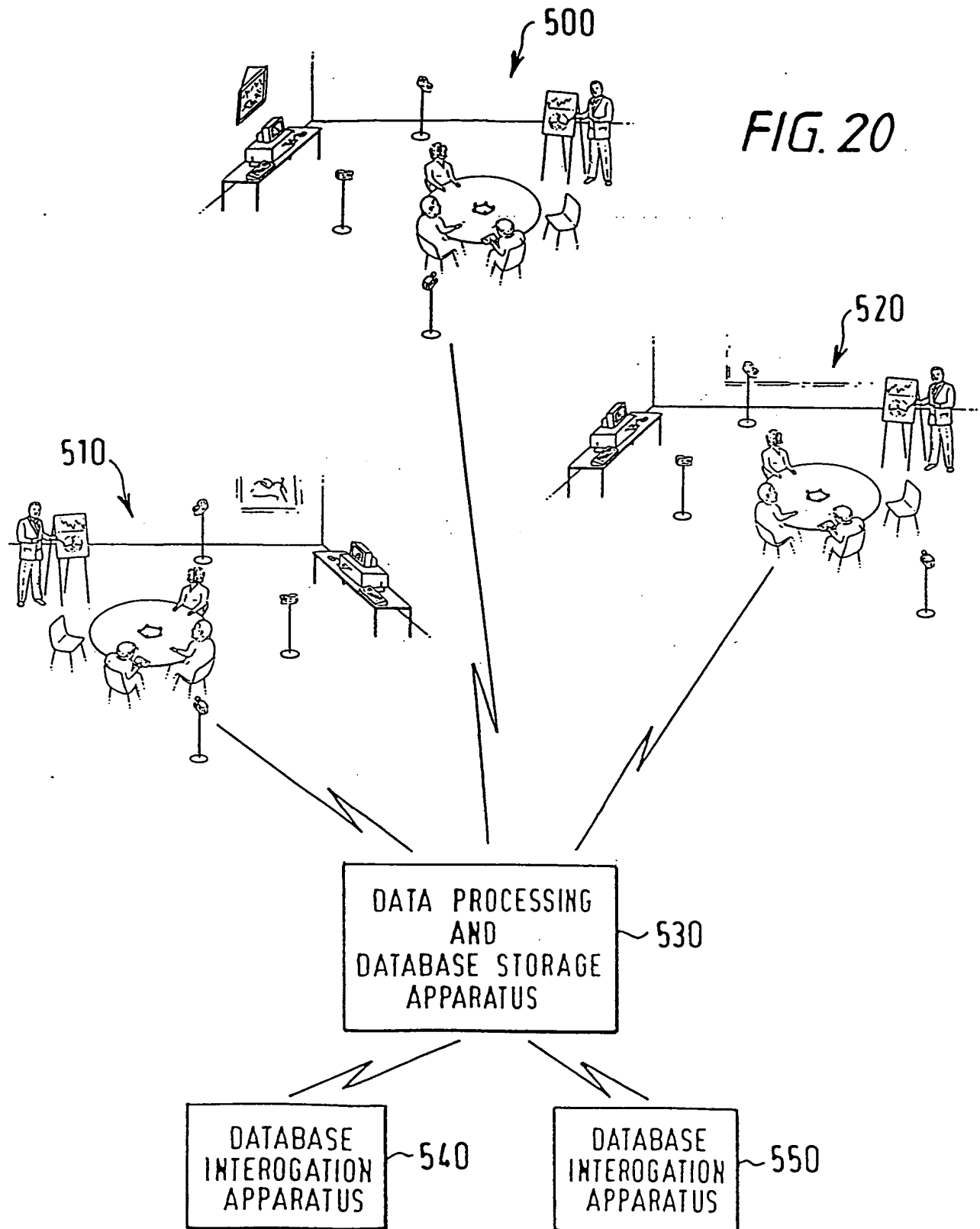
FIG. 19 A

The following parts of the meeting are relevant. Please select one for playback:

1. Speech starting at 10 mins 0 secs (0.4 x full meeting time)
2. Speech starting at 12 mins 30 secs (0.5 x full meeting time)

FIG. 19 B

THIS PAGE BLANK (USPTO)



THIS PAGE BLANK (USPTO)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☒ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☒ GRAY SCALE DOCUMENTS
- ☒ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

THIS PAGE BLANK (USPTO)